The Dissertation Committee for Mark Arjun Bond certifies that this is the approved version of the following dissertation:

Using IRT parameters as informative priors in second-order Bayesian latent growth modeling

Committee:

Susan N. Beretvas, Supervisor

Tiffany A. Whittaker

Keenan A. Pituch

Carlos M. Carvalho

Jodi Casabianca

Using IRT parameters as informative priors in second-order Bayesian latent growth modeling

by

Mark Arjun Bond

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2018

Acknowledgements

First and foremost, I would like to thank the Ethiopian goat herder who discovered coffee. If you had not noticed that your flock became energetic after eating the fruit, our world would be very different. Without your careful attention, this dissertation, and indeed my entire scientific career would be seriously degraded. Just as Newton stood on the shoulders of giants, so do I stand on the shoulders of goats.

Special thanks are also due to Dr. Beretvas for extremely thorough support throughout this process, and Dr. Carvalho for support developing the mathematics. I appreciate the time every committee member spent on this work. You are all very busy people. Thanks also to Dr. Cawthon for giving me practical experience with longitudinal data. Without that experience, undoubtedly this more theoretical work would suffer.

Finally, thanks are in order to Dr. Bond for careful line-editing of the manuscript, and to Dr. Pitre for introducing me to Hamilton (1994), a pivotal text in this area. I note with amusement that an unfamiliar reader might mistake these individuals for faculty members.

Using IRT parameters as informative priors in second-order Bayesian latent growth modeling

by

Mark Arjun Bond

The University of Texas at Austin, 2018

Supervisor: Susan N. Beretvas

In education, a wide variety of statistical methodologies are available to study change over time. For example, second-order latent growth models correct for item characteristics while estimating student-level growth. However, second-order latent growth models are difficult to estimate, with low convergence rates and high bias (Murphy, Beretvas, and Pituch, 2011). In attempting to correct this, I proposed and evaluated a new estimation method using the Kalman filter and informative priors for item parameters. This fully Bayesian estimation method was theoretically guaranteed to converge eventually, while informative parameters, theoretically justified within Item Response Theory, were hypothesized to reduce the mean squared error of parameter estimates. However, a simulation study found several scaling problems with the estimation method, and estimation using a real data set failed to converge. Discussion provides a few recommendations to correct these scaling problems.

Table of Contents

List of Tables

List of Figures

Introduction

Social scientists are often interested in understanding change over time. Within the context of educational psychology, the process of learning is one form of change over time. When student learning is carefully measured and properly modeled, it is possible to address a wide variety of research questions. On the most basic level, is the learning process exponential, quadratic, or linear in nature? More practically speaking, why do some students learn more than others? What can be changed to help students achieve?

Researchers can investigate these questions with latent growth modeling, a particular form of structural equation modeling (Bollen & Curran, 2006). In the simplest latent growth model, observed scores are modeled as a function of a random intercept term and a random slope term. In other words, the model can be used to provide an initial ability and growth term estimate for each person. Many of the above questions can be answered within this basic framework by adding explanatory variables to the growth term.

Recent research has focused more on what are called second-order latent growth models. In these models, multiple items are measured at each time point, and the measurement error in the items are separated from the true latent scores which are modeled as the direct indicators of the growth model's parameters. Simulation studies have indicated that second-order latent growth modeling has a number of advantages, including a greater probability to detect statistically significant differences between growth terms (von Oerzen, Hertzog, Lindenberger,

& Ghisletta, 2010), better estimates of item reliability (Geiser, Keller & Lockhart, 2013), and the ability to test whether the measurement model changes over time (Ferrer, Balluerka, & Widaman, 2008).

It is possible that, even with a measurement component, the basic framework of latent growth modeling as described may not adequately model change over time. Specifically, this form of latent growth modeling does not take into account how observations that are adjacent in time may be related to one another. It is possible that achievement at a particular time point might be correlated with adjacent achievement scores above and beyond the association modeled between initial ability and student growth. Simulation studies have suggested that failing to correctly model autocorrelation over and above the associations captured by the linear growth parameter can negatively influence results by resulting in greater type 1 error rates when testing whether the growth term is nonzero, along with reducing power for detecting nonzero growth (Murphy & Pituch, 2009; Kwok, West, & Green, 2007).

Most recently, Murphy, Beretvas, and Pituch (2013) conducted a simulation study evaluating the estimation of an autoregressive second-order latent growth model. Findings indicated that although modeling an autoregressive trend reduced relative bias, variance terms remained substantially biased. Furthermore, when an autoregressive model was estimated, between 14.3 and 43.7% of the models failed to converge.

To attempt to address these limitations, the present work seeks to expand upon Murphy et al., (2011) in a number of ways. First, a Bayesian estimation method is proposed. In this estimation method, a particular Gibbs sampler, the Kalman filter, accounts for the autocorrelated nature of the indicators of the latent constructs. Unlike the methods employed in Murphy et al. (2011), this Gibbs sampler is theoretically guaranteed to converge eventually (Roberts & Smith, 1994). Secondly, informative priors for measurement model parameters for the items are proposed and justified. Finally, a simulation study is proposed and conducted wherein the use of informative priors for measurement is evaluated, along with a real data analysis which uses the proposed Bayesian estimation method.

Literature review

This literature review will cover six main topics. First, the mathematical and statistical specification of first-order latent growth modeling is introduced. The next section presents the second-order latent growth model along with results from simulation studies that have assessed estimation of the model. A variety of time series approaches will be reviewed. Then Bayesian methods for time series analysis, particularly the Kalman filter, will be introduced. The justifiability of using informative priors in the context of this research will be discussed, along with a brief discussion of calibrating Item Response Theory models. Finally, relevant literature for a real data analysis will be reviewed.

First Order latent growth modeling

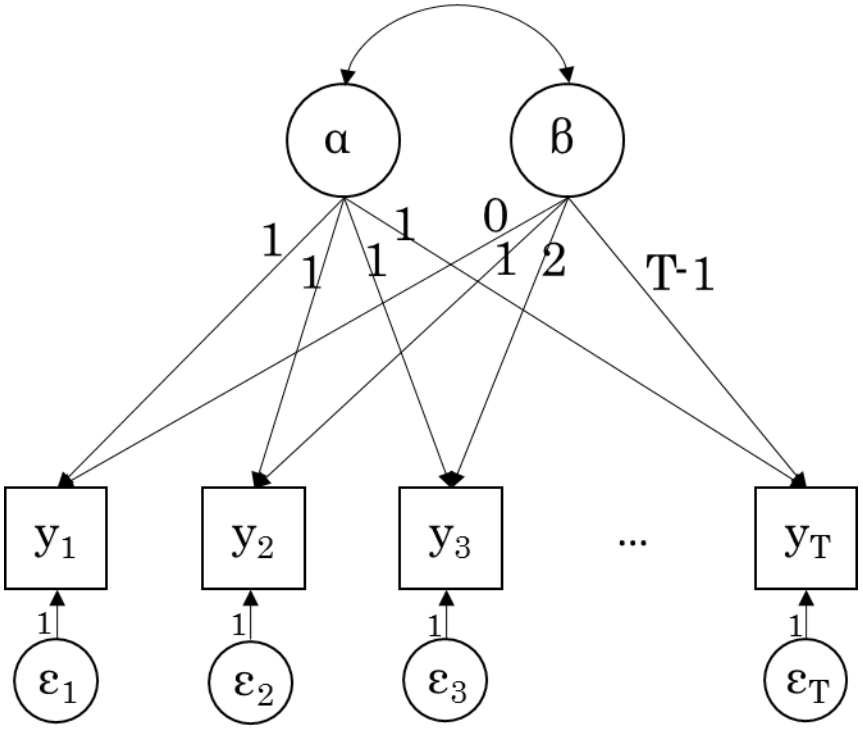
In educational psychology, it sometimes happens that large-scale testing data are collected longitudinally so that researchers can evaluate change over time. Researchers may then wish to assess the impact of various explanatory variables on student growth in achievement outcomes. Although growth over time may be analyzed using analysis of variance (ANOVA) or conventional multiple regression, doing so assumes that all students grow at the same rate. Another form of analysis, latent growth models using structural equation modeling (SEM) or multilevel modeling allows growth trajectories to vary by student (see, for example, Hancock, Harring, & Lawrence, 2013). Using the SEM framework, the model is mathematically specified in the following way (Bollen & Curran, 2006):

$$y = \boldsymbol{\Lambda}\eta + \varepsilon \qquad (1)$$

where y is the $T \times 1$ vector of observed responses for student i, η is the $m \times 1$ vector of m factors for student i, Λ is the $T \times m$ matrix of factor loadings, and ε is the $T \times 1$ vector of residuals, which are typically assumed i.i.d. Normal with a mean of zero.

Supposing that growth in a latent construct is hypothesized to be parameterized using only an intercept term and a linear growth term with equally spaced time intervals, m equals two, and the first column of Λ contains the value *1*, whereas the second column contains the time variable coded as follows $t = (0, 1, 2, \ldots, T - 1)^t$, where ^{t} denotes the transpose of that row vector. Also in this example, we denote the first element of η as α and the second element as ß, which are, respectively, the intercept and growth terms. A structural-regression diagram (explained in, e.g., Kline, 2011) for this running example is provided in Figure 1, below. The model is defined as "unidimensional" since α and ß denote the intercept and slope of a single latent factor; it is defined as "first-order" since latent measurements at each time point are unmodeled.

Figure 1. A structural-regression diagram of a linear unconditional unidimensional first-order latent growth model



The observation, y_t*, for each time point* t *is related to the covarying random intercept and slope factors α and ß. Observations at each time point have a separate random error term,* ε_t.

Next, consider how exactly α and ß in η are specified as being random. They may be expressed as function of an average and a deviation from the average:

$$\eta = \mu_\eta + \zeta \qquad (2)$$

where the $m \times 1$ mean vector μ_η contains elements μ_α and $\mu_ß$, and the $m \times 1$ residual vector ζ containing ζ_α and $\zeta_{ß,}$ which are assumed to be normally distributed with

means of zero. It is assumed that ζ and ε are independent of each other, however, the elements of ζ are not typically assumed to be independent of each other, so their $m \times m$ covariance matrix $\boldsymbol{\Psi}$ is given by:

$$\boldsymbol{\Psi} = \begin{bmatrix} \psi_{\alpha\alpha} & \psi_{\alpha\beta} \\ \psi_{\alpha\beta} & \psi_{\beta\beta} \end{bmatrix} \qquad (3)$$

where variance terms, $\psi_{\alpha\alpha}$ and $\psi_{\beta\beta}$, correspond to the intercept and slope terms, respectively, and $\psi_{\alpha\beta}$ represents the covariance between the intercept and slope factors.

The covariance matrix of the latent factors may be used to compute the variance of the observed responses. Plugging Equation 2 into Equation 1 gives:

$$y = \Lambda(\mu_\eta + \zeta) + \varepsilon \qquad (4)$$

meaning that the model-implied covariance matrix $\sum$ can be expressed as

$$\sum = \boldsymbol{\Lambda\Psi\Lambda}^t + \boldsymbol{\Theta}\varepsilon \qquad (5)$$

where $\boldsymbol{\Theta}\varepsilon$ is a $T \times T$ diagonal matrix of variance terms for ε at each time point.

An HLM specification of the latent growth model is theoretically identical to the SEM specification (Curran, 2003). In fact, structural equation modeling subsumes certain kinds of hierarchical linear modeling in the sense that many different kinds of multi-level structures may be represented in an SEM (e.g., Muthén, & Asparouhov, 2011, Curran, 2003). However, a unidimensional linear latent growth model may be estimated in SEM using only a covariance matrix and a

mean vector, whereas estimating these same models using conventional HLM software requires the raw data (Curran, 2003). Beyond concerns about convenience, it is more conventional in the literature to use SEM to model measurement error (contrast, e.g., Schumacker & Lomax, 2010, with Gelman & Hill, 2007). Since this project ultimately aims to incorporate a measurement-error model, the SEM notation will be used.

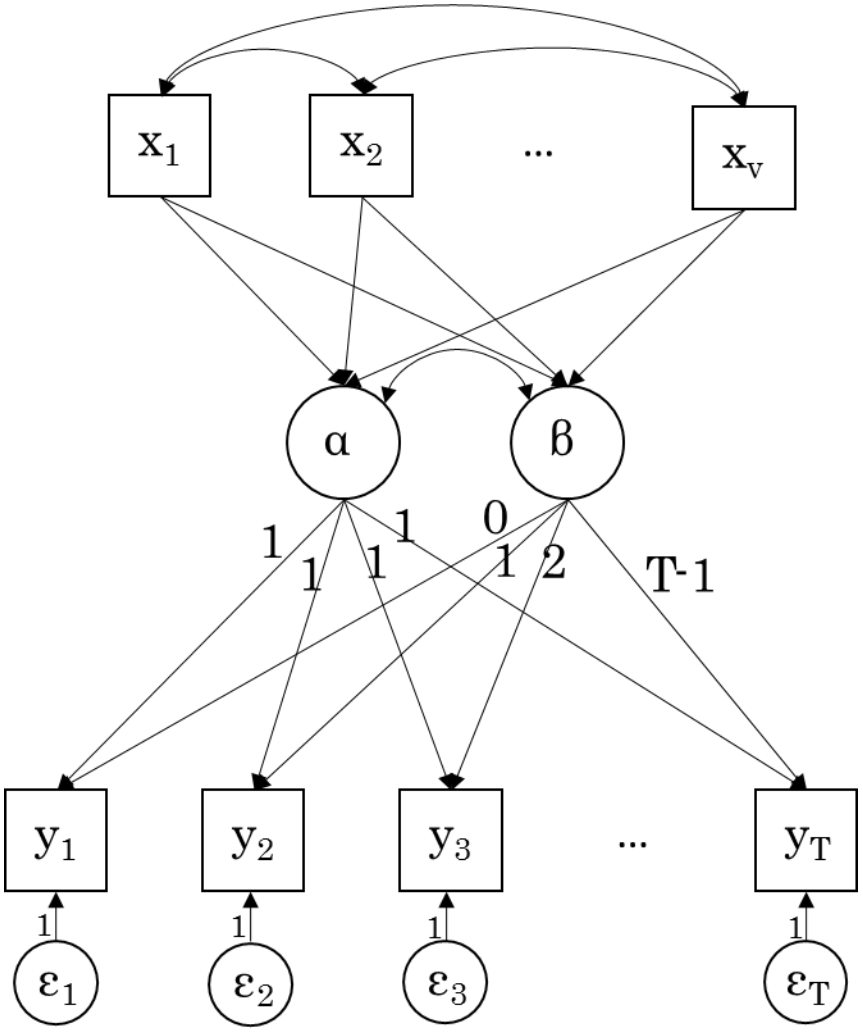
Regardless of the parameterization employed, an unconditional LGM may be of somewhat limited use to social science researchers. That is, an unconditional LGM merely describes growth in the relevant variables and variances in the growth trajectory parameters. Under the definition of "unconditional", there are no predictor variables modeled as explaining variability in growth parameters. Researchers may be more interested in the reasons why, for example, certain students learn more quickly than others, rather than a description of students' average growth. To explain the variance in the latent intercept and growth terms, the conditional latent growth model is introduced.

Consider again a unidimensional linear latent growth model, this time with a $v \times 1$ vector of explanatory variables, x, with, for instance, variables that are called x_1, x_2, …, x_v. For the sake of argument, suppose that all v variables were collected prior to obtaining the first outcome measure. A structural-path model for such an example is shown in Figure 2, below. Then, keeping the notation the same as in the above example, once again, the observed outcomes are specified as in Equation 1, above. However, the latent vector η is modeled conditionally:

$$\eta = \mathbf{\Gamma} x + \zeta \qquad (6)$$

where $\mathbf{\Gamma}$ is an $m \times v$ matrix where each row represents regression weights for the m^{th} latent factor. If, as is typically the case, the variable x_1 contains a 1 for each student, then the first column of Γ contains the conditional intercept term, which, in this example, would represent the α and ß terms in Figure 2 when all the other elements of x are set to zero. Once again it is assumed that ζ and ε are independent of each other; additionally, it is now assumed that ζ and x are independent of each other. Again, the covariance matrix $\mathbf{\Psi}$ for the elements of ζ need not be diagonal, and is specified in Equation 3. Of course, in a conditional model, the elements of $\mathbf{\Psi}$ represent the variances and covariances that are not explained by the terms in x.

Figure 2. A structural-regression diagram of a linear unidimensional first-order latent growth model, where the intercept and slope are conditional on predictors x_1 through x_v.



Recall that, in this example, the explanatory variables x_1, x_2, …, x_v were assumed to have been collected prior to testing. Also, the growth term ß remains the same at each time point for a given participant. It may be that researchers are interested in a more dynamic system in which, for instance, growth varies from

time point to time point, as a result of changing independent variables. A more dynamic model would be more appropriate when a great deal of time-sensitive data are collected, and when it is straightforward to determine which variables are independent of the measurement process. However, for the sake of simplicity, the present work focuses on growth models in which the growth term is fixed for a particular person.

Once a conditional model is fit, researchers may evaluate the matrix $\mathbf{\Gamma}$ to assess the degree to which particular predictor variables are related to initial student ability and student growth. Whereas unconditional latent growth modeling merely provides a description of student growth, it is possible for conditional latent growth modeling to regress those growth terms on student characteristics.

Both of the above models, which were diagrammed in Figures 1 and 2, may be limited since a linear growth term is assumed. However, nonlinear terms may be easily incorporated. Suppose that researchers wish to estimate an intercept, a linear growth term, and a quadratic growth term. Then, η may be expanded to include α, ß, and an extra term, say, δ, meaning that m, the number of dimensions describing the trajectory, is equal to three instead of two. One may then specify the observed outcomes as in Equation 1, above, with the third column of $\mathbf{\Lambda}$ written as the vector $(0^2, 1^2, 2^2, \ldots, T-1^2)^t$. In other words, to estimate a squared growth term, researchers can just square each element of t, (which was defined under Equation 1), and add another element of η. Any polynomial growth term may be specified in this way, provided that there are enough time points for that polynomial to be

identified (Bollen & Curran, 2006). In a similar fashion, logarithmic or exponential growth terms may be specified by applying analogously appropriate functions to each element of t (e.g., Grimm, Ram, & Hamagami, 2011; Grimm & Ram, 2009). Beyond the augmentation of η with additional terms and that of $\mathbf{\Lambda}$ with additional columns, both conditional and unconditional nonlinear latent growth models may be specified using the same equations as above.

Even with explanatory variables and nonlinear growth terms, latent growth modeling as described above is limited in several ways. First, note that the observed responses for each student at each time point are modeled as a single summary measure. In other words, in this simplest growth model, a total or average score is modeled for the outcome rather than item score by item score. Thus, the measurement error of each specific item is subsumed into a single error term for an observed outcome. Also, another kind of error, which could be present, is unmodeled in a first-order latent growth model. Possibly, a student's latent ability at time t could depend on circumstances that are particular to the time point t. For example, a measurement of ability at a specific time could be error-prone due to a flawed test administration procedure during that time. In other words, there may be time-specific error sources for each ability estimate that could impact the estimation of student ability which are unrelated to measurement error from the items. However, in first-order latent growth modeling, time-specific error and measurement error are combined into one term. Properly separating these terms would allow a more accurate apportioning of measurement error.

Another potential limitation arises from the way that these models are typically estimated. First-order latent growth models may be estimated using maximum likelihood-based estimation, relying on iteratively updating proposed model estimates and checking a discrepancy function (Bollen & Curran, 2006, pp. 39-42). Unfortunately, this local discrepancy function may achieve a local minimum; also, the process of iteratively updating a likelihood-based estimate may fail to converge. Latent growth models may also be estimated using Bayesian methods, such as a Gibbs sampler (e.g., Elliott, Gallo, Ten Have, Bogner, & Katz, 2005), which are mathematically guaranteed to converge eventually (Roberts & Smith, 1994). Gibbs samplers are discussed more thoroughly in the section describing the Kalman filter. However, researchers have used a likelihood-based methodology in a wide variety of applications, which are discussed in the next section.

A great deal of research has been done using structural equation models that are similar to the one diagrammed in Figure 2, above. That is, although researchers often incorporate nonlinear techniques in fitting a latent growth model, it is often the case that researchers do not estimate a latent ability term for every student at each time point. For example, one study used a unidimensional latent growth model to estimate the effect of parental support and monitoring on adolescent alcohol use (Barnes, Reifman, Farrell, & Dintcheff, 2000). The two latent variables, as shown in Figure 2, were baseline alcohol misuse (analogous to α) and the slope in this term (analogous to β). Notably, the vector y in this article was a composite measure, taking the average of three separate item scores. Then, the measurement error ε in

some way represents the measurement error of this composite term. Possibly it would be of use to estimate the measurement error of specific items also, since researchers could then assess whether those items are worth including. This composite error term also confounds measurement error with time effects; a particular time point may have more or less error for reasons beyond measurement error. For instance, in the first wave of data collection, the participants were between 13 to 16 years of age, and in the last wave of data collection, participants were between 18 to 22 years old. Because some of the students may have not been living with their parents in the last wave, one might argue that some participants had a greater incentive to lie about their alcohol use in the first wave, relative to the last wave. This is error from an effect that may change over time, that is, error due to a time effect. This is theoretically distinct from those measurement errors that are not associated with a time component. However, in Barnes et al. (2000), those two effects are confounded in the ε term. Moreover, it is assumed that the measurement error has the same variance across all time points, which may or may not be realistic, depending on the topic of study. Finally, when items are parceled, it is assumed that the averaged items are unidimensional and that they are all equally indicative of the latent construct (Little, Rhemtulla, Gibson, & Schoemann, 2013). If the items are not actually unidimensional, there is a threat to construct validity which cannot be evaluated.

To summarize, when researchers fail to specify an exact measurement model for the observed outcomes, (the *y*s), measurement error and time-specific error are

confounded. Measurement error is assumed to be homoscedastic, which may or may not be appropriate, depending on the area of study. Also, it is assumed that the items are unidimensional. Unfortunately, a good amount of applied research has been done in this way, where the observed outcome is some composite measure, and the psychometric properties of the individual items are not statistically modeled (see, e.g., Fleming, Mason, Mazza, Abbott, & Catalano, 2008; Colder et al., 2001; Barnes et. al, 2000; Chan & Schmitt, 2000; Duncan, Duncan, Biglan, & Ary, 1998). All this research may have been better served to model the measurement error of each item separately, with each item loading onto a latent measure of ability at each time point. That is the approach of what is called second-order latent growth modeling.

Second-Order Latent Growth Modeling

In second-order latent growth modeling, there is a latent measure of a trait for each participant at each time point, modeling the measurement error of each item individually (See, e.g., Harring, Kohli, Silverman, & Speece, 2012; Sayer & Cumsille, 2001). Suppose that, for each time point, every participant responds to K items. Then let y be the $(K * T) \times 1$ vector with elements defined so that the first element is the response to first item at the first time point, the second element is the response to the second item at the first time point, and so on. Call the elements of this vector y_{11}, y_{12}, …, y_{1K}, y_{2K}, …, and y_{TK}. Also, let η be a $(T * m) \times 1$ vector, whose first element is the value of the first latent variable at the first time point, whose second element is the value of the second latent variable at the first time

point, and so on; call the elements of this vector η_{11}, η_{12}, …, η_{1m}, η_{21}, …, and η_{Tm} Furthermore, let η_1 be the vector formed by the first m elements of η, let η_2 be the vector formed by the next m elements of η, and so on. Note that the quantity T*m represents the total number of latent constructs at level one. Then the measurement model may be written identically to Equation 1, but with $\mathbf{\Lambda}$ written as a $(K * T) \times (T * m)$ matrix and ε written as a $(K * T) \times 1$ vector instead. Similarly to y, above, call the elements of ε ε_{11}, ε_{12}, …, ε_{1K}, ε_{2K}, …, and ε_{TK}. Each row of $\mathbf{\Lambda}$ gives the factor loadings for a particular item at a particular time point. The matrix $\mathbf{\Lambda}$ is assumed to be block-diagonal; in other words, trait estimates at a particular time point do not load onto items at another time point. If it is assumed that the measurement model is the same across time points, then all the blocks that make up $\mathbf{\Lambda}$ are equal to the $K \times m$ submatrix $\mathbf{\Lambda_a}$, formed by taking only the first K rows and first m columns of $\mathbf{\Lambda}$. Supposing that the measurement equation is written in this way, the structural portion may be written similarly to Equation 6 (Hancock, Kuo, & Lawrence, 2001):

$$\eta = \mathbf{\Gamma}\xi + \zeta \qquad (7)$$

with ζ an $(T * m_2) \times 1$ error vector that is similar to the same term in Equation 6, with $\mathbf{\Gamma}$ an $(T * m) \times m_2$ matrix of factor loadings, and with ξ an $m_2 \times 1$ vector of second-order latent factors; note that the number of second order latent factors, m_2, need not be the same as the number of first-order latent factors, m. Typically, models estimate fewer second-order factors than first-order factors (see, e.g., Hancfock, et al., 2001). Similarly to η, call the elements of ζ ζ_{11}, ζ_{12}, …, ζ_{1T}, ζ_{21}, …,

and $\zeta_{m_2 T}$; furthermore, let ζ_1 be the vector formed by the first T elements of ζ, let ζ_2 be the vector formed by the next T elements of ζ, and so on. Finally, the vector ξ is specified similarly to Equation 2:

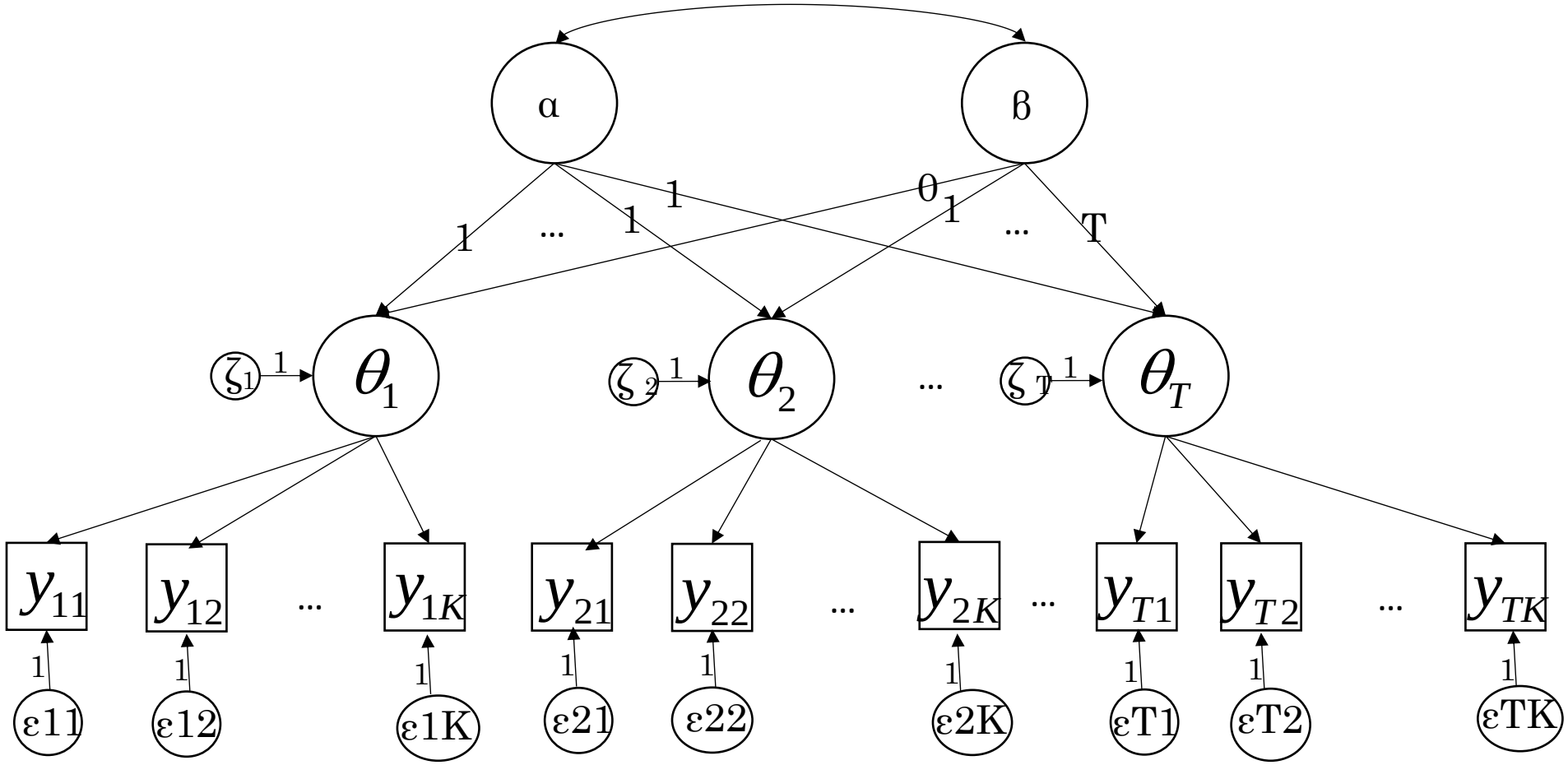
$$\xi = \mu_\eta + \zeta_\eta \qquad (8)$$

with μ_η an $m_2 \times 1$ mean vector and with ζ_η the $m_2 \times 1$ disturbance vector representing the variance of the level-two latent factors, also assumed to normally distributed with means of zero.

For the sake of example, consider a linear unconditional unidimensional second-order latent growth model, assuming strong metric invariance across time points. A structural-regression diagram for such a model provided in Figure 3, below. Begin by considering Equation 8 in the context of this example. Since the model is assumed to be linear, unconditional, and unidimensional, let $m_2 = 3$, and allow the mean vector to consist of the elements 1, μ_α, and μ_β; then, let the elements of ζ_η be 0, $\zeta_{\eta\alpha}$ and $\zeta_{\eta\beta}$. The quantities 1 and 0 in these vectors are set as fixed. Then, the vector ξ consists of the elements 1, α, and β; the latter two are, respectively, intercept and growth terms. Then, in Equation 7, the matrix $\mathbf{\Gamma}$ is block-diagonal, with blocks *1*, *1*, and *t*, as defined in Equation 1. Computing the matrix algebra, $\eta_1 = 1$, and $\eta_2 = \alpha * 1 + \beta * t$; then, $\zeta_1 = 0$. For convenience, call the elements of η_2 $\theta_1, \theta_2, \ldots, \theta_T$, and call the elements of ζ_2 $\zeta_1, \zeta_2, \ldots, \zeta_T$. Next, consider the measurement model as defined in Equation 1. Since strong metric invariance is assumed, it suffices to consider the $K \times m$ submatrix $\mathbf{\Lambda_\alpha}$. Call the respective columns of this matrix y_1 and y_2; call the

elements of these column vectors γ_{11}, γ_{12}, …, γ_{1K}, and γ_{21}, γ_{22}, …, and γ_{2K}, respectively. Computing the block-matrix algebra, then, the product vector on the right hand side of Equation 1 is equal to $\gamma_1 + \gamma_2 * \eta_2$.

Figure 3. A structural-regression diagram of an unconditional linear unidimensional second-order latent growth model, assuming strong metric invariance across time points.



Recall that, typically, the error terms ε are assumed to be independent identically distributed Normal variables. If the vector y represents binary data, for example, if each element of y records whether or not an examinee correctly answered a question, then the assumption of normality would not be appropriate. Whereas normally distributed variables have support along the entire real line, in this case each element of ε would be restricted to lie between negative one and positive one. To avoid this inappropriate specification, a different link function for the vector y may be employed:

$$P(y = 1) = \Phi(\gamma_1 + \gamma_2 * \eta_2) \qquad (9)$$

with $\Phi(x)$ representing the cumulative density function of the standard normal distribution. This specification may be thought of as an implementation of the Normal Ogive model in psychometrics (Lord, 1952; see also Fox & Glas, 2001).

Regardless of whether the vector y represents continuous or binary data, the second-order latent growth model has more variance terms than a first-order latent growth model. The error terms ζ_2 and ζ_η theoretically represent distinct forms of variation (Sayer & Cumsille, 2001). That is, the terms in ζ_η represent variation in the initial state and the variation in the growth term; the terms in ζ_2 represent unaccounted variation in the latent factors θ. Also, for continuous y vectors, the terms in ε represent unaccounted variation in the vector of observed variables y. In other words, the vector ζ_2 represents how the latent factor itself varies over time, independently of measurement error. In contrast, consider vector ε represented in Figure 1.

For the sake of example, consider a latent growth model estimating the growth of mathematics achievement in grade-school children, assuming the data are Normally distributed. Suppose that, for the first testing occasion, the children happened to have unusually variable achievement, that is, achievement that is more variable than is observed at later time points. Suppose also that the questions asked in the first assessment were unusually poor at discriminating between high and low achieving students. In a first-order latent growth model, these two sources of variation, that is, the unaccounted variation in mathematics achievement over time and the variation due to measurement error, would both have to be contained in a single error term, ε_1, since that is the only possible random error associated with the first measurement, y_1. In this way, these two sources of variation are confounded in a single error term (Geiser, Keller, & Lockhart, 2013).

In contrast, a second-order latent growth model would not have this difficulty. Heteroskedastic mathematics scores would be captured in the variance of the term $\zeta_{I,}$ while poor measurement during the first testing would be modeled in the variance of the error terms ($\varepsilon_{11,}$ $\varepsilon_{12,}$ …, and ε_{1K}).

The direct measurement of unaccounted (or time-specific) factor variance allows for further statistical estimation and testing. For instance, consider the fitting of unconditional and conditional second-order latent growth models. Researchers may be interested in the proportion of variation in latent factors that is accounted for by conditioning those factors on exogenous variables. In second-order latent growth models, but not first-order latent growth models, this may be

accomplished by comparing the unexplained variance of the latent factors before and after these exogenous variables are included in the model. Moreover, in second-order latent growth models, researchers may test the null hypothesis that the unaccounted variation in the factors is equal to zero. A statistically significant result of this test would indicate that more variables may be necessary to fully account for the time-specific variation in the constructs. Again, this statistical test would not be possible for a first-order latent growth model, since factor variation is not explicitly included in that model.

Moreover, if this time-specific variance exists, then the reliability of the items is better estimated in a second-order latent growth model. Reliability in this context is defined as the variance of the latent variable divided by the total observed variance (Steyer, Ferring, & Schmitt, 1992). In Geiser, Keller and Lockharts' (2013) simulation study, it was shown mathematically that first-order latent growth models would, on average, tend to underrepresent the reliability of the items, given the existence of time-specific variance. Next, in that 2013 simulation study, differing amounts of occasion-specific variance were used to generate latent growth curves, then reliability was estimated using both first-order and second-order latent growth models. Finally, percentage bias terms were obtained by taking the difference of estimated and true reliability terms, dividing by the true reliability. When there was no time-specific variance, the first-order and second-order latent growth models were roughly comparable in terms of percentage bias. However, if there was even a small amount of occasion-specific variance, the first-order latent

growth model had between 12.1 to 29.9 percent bias, whereas the second-order latent growth model had only between 0.01 to 0.07 percent bias. This trend increased as occasion-specific variance increased, with a maximum percentage bias of 64.9 percent for first-order latent growth models, and a maximum percentage bias of 12.0 percent for second-order latent growth models. Thus, when there is a great deal of time-specific variance, there is a strong incentive to use second-order latent growth modeling rather than first-order latent growth modeling.

Suppose alternatively that there is relatively little time-specific variance when modeling latent growth. In this situation, it has been shown analytically (von Oerzen, Hertzog, Lindenberger, & Ghisletta, 2010) that second-order latent growth models have greater power to detect individual differences in growth than first-order latent growth models. In a latent growth modeling context, the power to detect individual differences in growth is the probability of correctly rejecting the null hypothesis that the variance in growth is equal to zero. Therefore, if the variability of this slope term is of substantive interest, a second-order latent growth model may be preferred. Moreover, a simulation study (von Oerzen et al., 2010) demonstrated that a larger number of observed variables increases the power to detect variability in intercept and slope terms, particularly when there is little time-specific variance. Since first-order latent growth models are unable to accommodate multiple observed variables at a single time point, the 2010 simulation study also provides evidence that second-order latent growth modeling is preferred when testing the variability of slope and intercept terms.

Second-order latent growth models may also allow for the factor loadings in Λ to differ, if needed. In other words, the measurement of the factor may be allowed to vary over time. It is also possible to statistically test for the need to do this (Ferrer, et al., 2008). Since the measurement of the factor is not explicitly modeled in first-order latent growth modeling, such a procedure would not be possible using a first-order latent growth modeling. Indeed, first-order latent growth modeling must assume that the measurement of the factor remains stable over time (Ferrer et al., 2008). In an empirical research article, Ferrer, et. al (2008) showed that when these factor loadings actually differ, the results from first-order and second-order latent growth modeling provide widely disparate results for the growth term. In short, second-order latent growth modeling can test the assumption of equal factor measurement across time; when this assumption is violated and untested, the growth term may be biased. Again, unfortunately, first-order latent growth models are unable to be used to test this assumption, since they assume that factor measurement remains the same over time (Ferrer et al., 2008).

Second-order latent growth models are relatively recent to the educational psychology literature (Leite, 2007). Their specification may be intimidating to applied researchers, which may be one reason why methodologists note a paucity of second-order latent growth models in the educational psychology literature (Geiser, Keller, & Lockhart, 2013). However, similar models have a long history in the econometrics literature.

Time Series Approaches

Consider, for instance, the unidimensional autoregressive model with noise (Granger & Morris, 1976; further discussed in Box, Jenkins, & Reinsel, 2008; Chatfield, 2004; Soares, Gonclaves, & Gamerman, 2009; Petris, Petrone, & Campagnoli, 2009), which can be adapted to subsume a second-order latent growth model. Recall that the first level of a conditional first-order latent growth model is specified as

$$y = \mathbf{\Lambda}\eta + \varepsilon \qquad (10)$$

with y, $\mathbf{\Lambda}$, η and ε as specified in Equation 1, and that a conditional first-order latent growth model further models η as

$$\eta = \mathbf{\Gamma}x + \zeta \qquad (11)$$

with η, $\mathbf{\Gamma}$, x and ζ as specified in Equation 6. For a unidimensional autoregressive model with noise, index the vectors y, η, and ε by the time point t, so that the first level of the model is given by

$$y_t = \mathbf{\Lambda}\eta_t + \varepsilon_t \qquad (12)$$

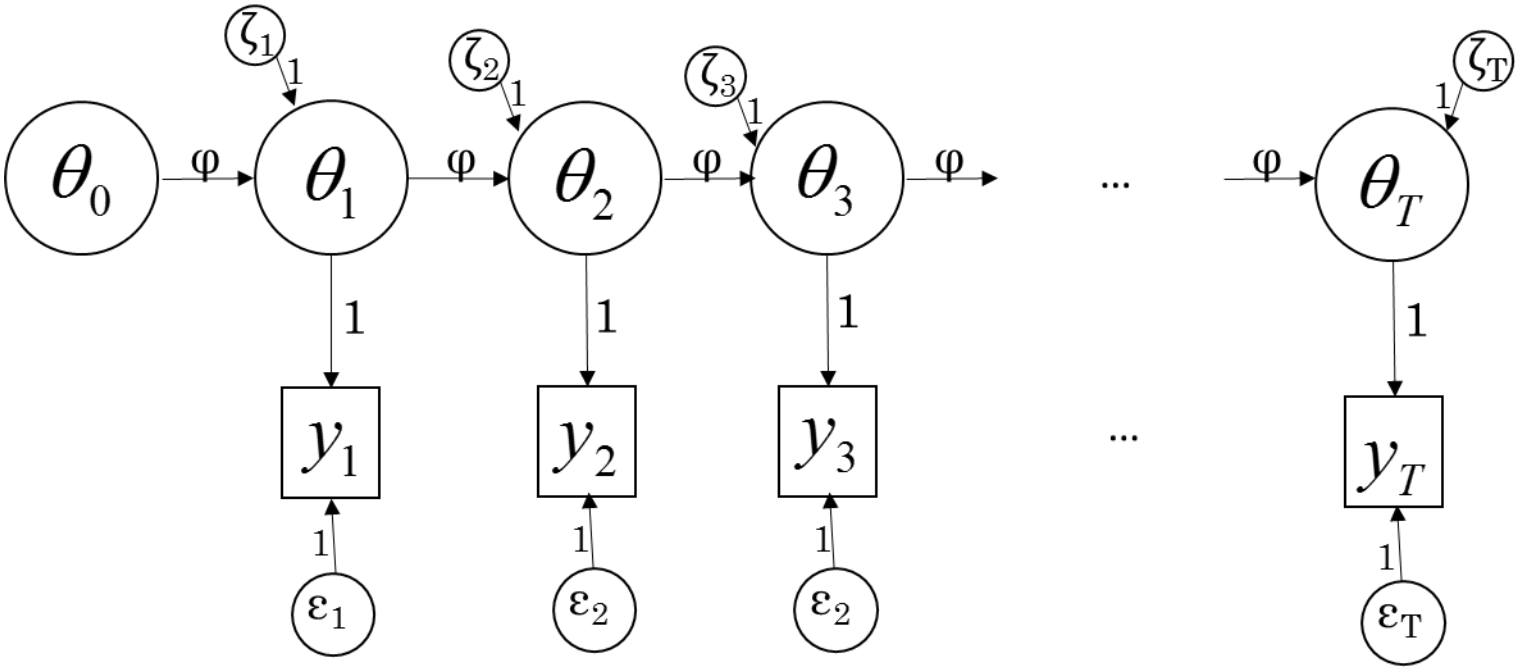
for time-indexed data vector y_t, time-indexed latent vector η_t and time-indexed error term ε_t. Since y_t is the vector of observed responses at time t, it is a $K \times 1$ vector, where K is the number of items. Then, $\mathbf{\Lambda}$ is a $K \times m$ factor loading matrix, where m is the number of elements in the column vector η_t, and, finally, let ε_t be a $K \times 1$ column vector. Then, the first column of the matrix $\mathbf{\Lambda}$ contains item parameters b_1, b_2, …, b_k, and the second column contains discrimination parameters a , a_2, …, a_k. Then, further model η_t as

$$\eta_t = \mathbf{\Gamma}\eta_{t-1} + \zeta_t \qquad (13)$$

with η_t and η_{t-1} as in Equation 12, above, with $\mathbf{\Gamma}$ as an $m \times m$ matrix of regression weights, and with ζ_t as an $m \times 1$ error vector. This is called an "autoregressive" or "AR" model since each construct is regressed on an earlier value of that same construct. Furthermore, this particular equation is called an AR(1) model, since η_t is only regressed on the one immediately previous value of η; it would also be possible, for instance, to regress η_t on the two most recent values, which would be called an AR(2) model.

For the sake of this example, let m equal one, and denote the element of η_t as θ_t. Then, the matrix $\mathbf{\Gamma}$ has a single element; call that element φ. The parameter φ is known as the "autoregressive parameter", since it is a regression weight describing the impact of θ_t on itself at a later time. A structural-regression diagram for this example is provided in Figure 4, below.

Figure 4. A structural-regression diagram of the unidimensional autoregressive time-series model with noise.



Note that Figure 4 includes an extra latent variable, namely, θ_0. This latent variable is included to ensure that the variance of the θ estimates remains constant over time. Without this latent variable, the variance of θ_1 would equal some value, say, ζ, and the variance of all future time points would equal $\varphi^2*\zeta$ (See, e.g., Hamilton, 1994 for the algebra confirming this). Therefore, the variance of these latent variables would differ across time points. Incorporating the latent variable θ_0, with variance ζ, ensures that the variance of all the other latent variables equals a fixed value over time. In particular, the variance of the θ estimates is equal to $\varphi^2*\zeta$ for each time point.

Like the first-order latent growth model, this autoregressive model may be modified in a variety of ways. For instance, rather than having a single observation at a time point, it is possible to include K observations y_{1t}, y_{2t}, …, y_{Kt} at each time point t, each with corresponding factor loadings γ_{21}, γ_{22}, …, γ_{2K}.

Employing this model necessitates the use of a number of assumptions. Similarly to the first-order latent growth model, in an autoregressive model it is typically assumed that the first order residuals ε_t are i.i.d., with means of zero, and are uncorrelated with each other at different time points (Hamilton, 1994). Also, it is typically assumed that the disturbances, ζ_t, are i.i.d., with means of zero, and are uncorrelated at different time points (Hamilton, 1994).

Additionally, it is commonly assumed that autoregressive time series models are at least weakly stationary. A time series process is said to be weakly stationary when two criteria are satisfied. The first criterion is that that the average value of the observed data remains the same across time points. The second criterion is that

the covariance between two observed values that are *j* time points apart is the same across time points, for any value *j*. In other words, a time series process is weakly stationary when the means, variances, and covariances of the observed data remains the same across time shifts, given the model (Hamilton, 1994). For this assumption to be met, any time-dependent trend in the data must be correctly modeled.

Weak stationarity is desirable for a variety of reasons. Firstly, long-term averages across all time points are mathematically guaranteed to converge to the population mean for autoregressive time series models (Hamilton, 1994). Also, given weak stationarity, it is possible to forecast an autoregressive time series model arbitrarily far into the future (Hamilton, 1994), allowing researchers to predict how the observed data may change in the future. Finally, assuming weak stationarity for autoregressive models allows for autocovariances to be obtained analytically (Hamilton, 1994). In a time series process, the j^{th} autocovariance is defined as the covariance between an observed datum point y_t and an observed datum point $y_{t-j,}$ for a particular value of *j*. The autocovariance function takes as input that number *j* and outputs the corresponding autocovariance for a particular time series. The autocorrelation and autocorrelation function are similarly defined, dividing the corresponding autocovariance by the stationary variance of the observed data.

To satisfy the assumption of weak stationarity, it is necessary, but not sufficient, for the autoregressive parameter φ in Figure 4 to be less than one in absolute value. To see this, consider the following autoregressive process without

noise. Let an observed datum point be a function of a previously observed datum point:

$$y_t = c + \varphi y_{t-1} + \varepsilon_t \qquad (14)$$

for y_t, y_{t-1}, φ, and ε_t as described above, and for c an intercept term. Suppose that this time series process is weakly stationary and recursively plug Equation 14 into itself. Then y_t may be expressed as

$$y_t = c + \varepsilon_t + \varphi(c + \varepsilon_{t-1}) + \varphi^2(c + \varepsilon_{t-2}) + \varphi^3(c + \varepsilon_{t-3}) + \cdots$$

$$= c + \sum_{i=0}^{T} \varphi^i \varepsilon_{t-i} + \sum_{i=0}^{T} \varphi^i \varepsilon_{t-i} \qquad (15)$$

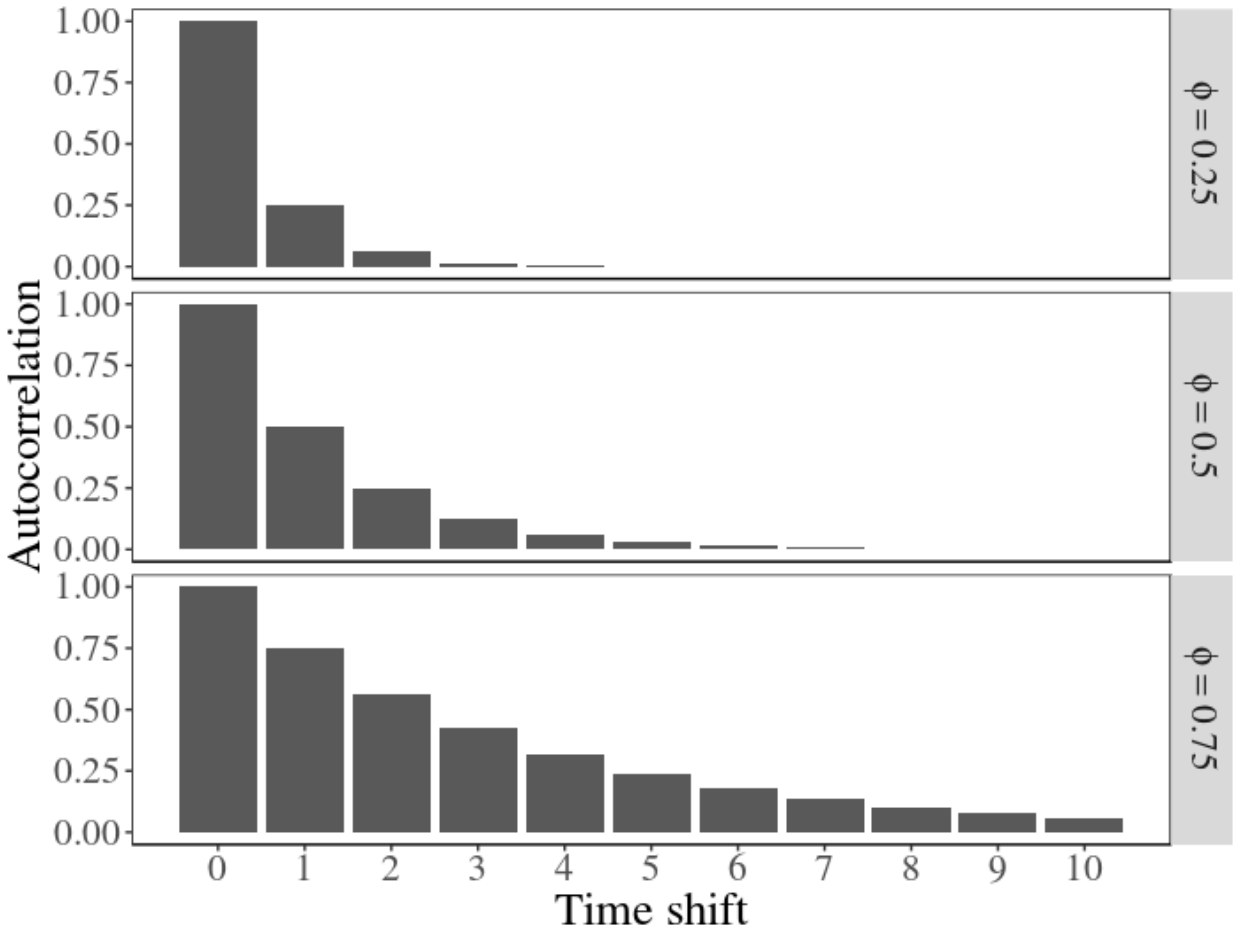
with all the above terms identical to Equation 14 above. Consider the limit as the term T becomes arbitrarily large. Supposing that φ is greater than one in absolute value, the term y_t will also increase in absolute value as the process continues, meaning that the average of the observed data would depend on the time point, violating weak stationarity. Suppose instead that φ is less than one in absolute value. Then, y_t may be expressed as

$$y_t = \frac{c}{1-\varphi} + \sum_{i=0}^{T} \varphi^i \varepsilon_{t-i} \qquad (16)$$

and the long-term expected value of the y_ts is $\frac{c}{1-\varphi}$, since the expected value of the error terms is assumed to be zero. Moreover, one may show analytically (e.g., Hamilton, 1994, p. 53) that the variances and covariances of this time series process remain constant across time points, satisfying weak stationarity. Specifically, the j^{th} autocorrelation is equal to φ^j, meaning that then autocorrelation function decays

geometrically over time. The autocorrelation function given different values of φ is plotted in Figure 5, below.

Figure 5. The autocorrelation function of an AR(1) process, given different values of the autoregressive parameter.



As indicated in Figure 5, in an AR(1) process, every observation is correlated with every other observation. However, depending on the value of the autoregressive parameter, this correlation may be quite small for data points that are far apart from each other in time. This sort of time series structure could be applicable to ability measurements in educational psychology, in which an ability measurement at a particular time point gives information about ability at a later time (see, e.g., Jordan, Kaplan, & Hanich, 2002, Catts, Bridges, Little, & Tomblin, 2008).

However, recall that the assumption of weak stationarity given the model must hold for such an autoregressive model to be employed. In particular, the assumption that the observed means remain the same over time given the model may be untenable without a growth term, particularly when student growth is the construct under study. It is therefore necessary to include some sort of growth term in an autoregressive model to properly apply it to educational psychology. In adding this growth term, a second-order autoregressive latent growth model will be outlined.

To introduce a growth term to the autoregressive unidimensional time series model, let the vector η_t in Equation 13 equal the column vector with elements 1, t, and θ_t. Then specify the matrix $\mathbf{\Gamma}$ as

$$\mathbf{\Gamma} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ \alpha & \beta & \varphi \end{bmatrix} \qquad (17)$$

with intercept term α, growth term β, and autoregressive parameter φ. Then, following Equation 13, η_{t+1} becomes:

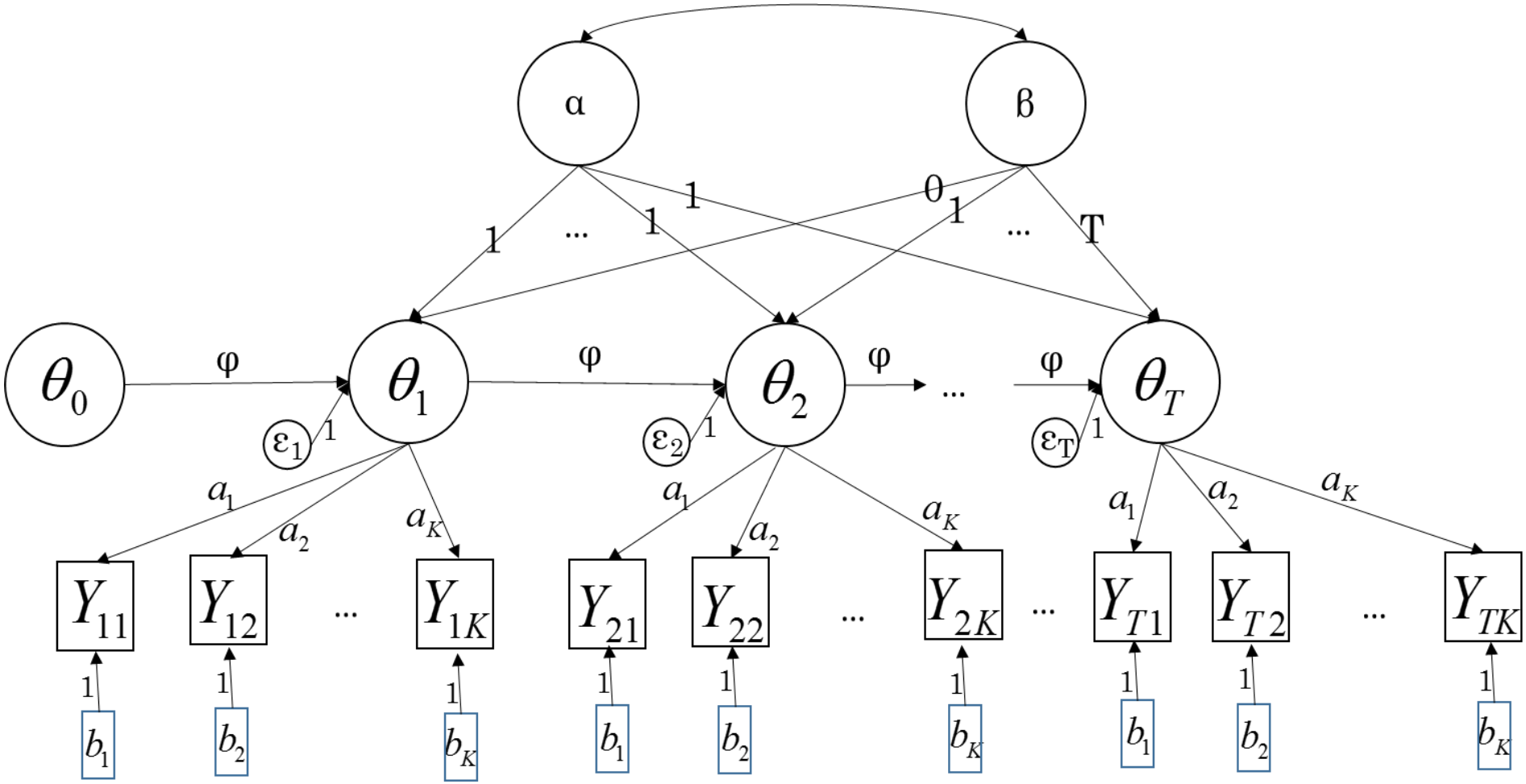
$$\begin{bmatrix} 1 \\ t+1 \\ \theta_{t+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ \alpha & \beta & \varphi \end{bmatrix} \begin{bmatrix} 1 \\ t \\ \theta_t \end{bmatrix} + \zeta_t \qquad (18)$$

with the first element of the error vector ζ_t set equal to zero. This specification is similar to the specification in Equation 14, however the first element of η_t may be used to generate a random intercept term in Equation 12, while the second element of η_t may be used to generate a random slope term in Equation 12. In other words, including the constants 1 and t in η_t allows for the estimation of individual-level growth and intercept terms. To allow these growth and intercept terms to vary, call the column vector consisting of the elements α, β, and φ ξ. Finally, let ξ be distributed as a multivariate Normal distribution, with notation similar to Equation 8:

$$\xi = \mu_\eta + \zeta_\eta \qquad (19)$$

with mean vector μ_η and disturbance vector ζ_η; call the variance-covariance matrix of this multivariate Normal distribution $\boldsymbol{\Psi}$, similarly to Equation 3, above. A second-order autoregressive latent growth model with this specification is shown in Figure 6, below.

Figure 6. A unidimensional second-order autoregressive latent growth model.



This second-order autoregressive latent growth model may be further modified or altered in many of the ways previously described in the latent growth modeling sections. For instance, researchers may be interested in including explanatory variables which may account for the variance in the intercept and growth terms of the model. Also, depending on the structure of the data, there may be many observed items collected at any particular time point, which may or may not be dichotomous. Again, supposing that the data are dichotomous, the link function shown in Equation 9 may be employed. In this way, all the advantages of second-order latent growth modeling may be retained while also including an autoregressive structure. Two measurements of student ability may remain significantly correlated even after correcting for that student's initial ability and growth, which would make the inclusion of an autoregressive term necessary. As discussed later, simulation studies indicate that failing to model an autoregressive trend when the data are generated with an autoregressive trend inflates type 1 error rates and biases and reduces power.

In simulation studies that model autoregressive processes, it is common to also model a moving average trend, another time series process (Hamilton, 1994). A moving average trend is defined by allowing previous error terms to have an impact on current observations:

$$y_t = c + \varepsilon_t + \theta\varepsilon_{t-1} \qquad (20)$$

where θ is the moving average parameter and y_t, c, ε_t, and ε_{t-1} are defined as in Equation 14. It may be shown analytically (Hamilton, 1994) that the first-order autocorrelation, that is, the correlation between the observations of two adjacent time points, is equal to $[\theta / (1 + \theta^2)]$. On the other hand, any higher-order autocorrelations are assumed equal to zero (Hamilton, 1994). In other words, the moving average time series structure allows observations that are adjacent in time to be correlated, while not forcing every point to be correlated with every other point.

However, as noted above, in educational psychology there are theoretical reasons to expect every observation to correlate with every other observation. More commonly, autoregressive and moving average trends are both modeled at the same time, in an autoregressive-moving average (ARMA) process:

$$y_t = c + \varphi y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1} \qquad (21)$$

with y_t, y_{t-1}, c, φ, ε_t, θ, and ε_{t-1} as defined in Equations 18 and 20. The specification in Equation 21 is called an ARMA(1, 1) process, since there is one autoregressive parameter and one moving average parameter. The autocorrelation function for an ARMA(1, 1) process is more complex mathematically (Hamilton, 1994), but it may be thought of as being similar to the autocorrelation function for an autoregressive process, with the first-order autocorrelation being altered by the moving average parameter.

When a time series trend is not adequately modeled by these autoregressive or moving average processes, researchers may instead decide to freely estimate each autocorrelation, rather than estimating them according to the autocorrelation functions provided above. This approach involves estimating more parameters than any of the previously discussed time series models. That is, for t time points, this would involve estimating $\frac{t(t-1)}{2}$ autocovariances (Kwok, West, & Green 2007). In contrast, the autocovariances for a first-order autoregressive process and a first-order moving-average process are completely specified by estimating a single parameter (Kwok, West, & Green, 2007). Moreover, the autocovariances of an ARMA(1, 1) process may be specified by estimating only two parameters, the θ and φ parameters.

There are a number of possible errors one might make when attempting to model a time series process. For instance, one might freely estimate all the autocovariances when an autoregressive trend would adequately model the data. Such a model would be over-parameterized. On the other hand, one might model all the observations as being independent when an autoregressive trend is actually present. This model would be under-parameterized. Finally, one might model an autoregressive trend as a moving average trend, or vice-versa, which would be generally mis-specified. Simulation studies have evaluated the practical impact that these errors have on estimation.

Simulation studies have indicated that failing to model an autoregressive term when such a trend actually exists may bias the estimation of a number of relevant

parameters for latent growth modeling. In a simulation study, Murphy and Pituch (2009) assessed the parameter recovery of a first-order latent growth model when time series trends were present. Failing to model these time series trends resulted in an elevated type 1 error rate when testing fixed effects. That is, for conditions with smaller samples, (N = 30), failing to model a time series trend resulted in type 1 error rates ranging from 0.079 to 0.125, when the nominal type 1 error rate was 0.05. This means that failing to include a time series component may lead researchers to incorrectly state that a variable explains student growth, or that a variable explains the students' initial ability.

More relevantly, Kwok, West, and Green (2007) simulated a condition in which data were generated with an autoregressive trend, but estimated without a time series trend. That is, longitudinal data were generated so that previous constructs impacted future constructs as in Equation 14, but the constructs were estimated as second-order latent growth models without including an autoregressive parameter. Then, relative bias of standard error and variance terms were presented, along with estimation of statistical power. When the estimation model was underspecified compared to the way the data were generated, the power to detect a nonzero intercept or slope was 0.28, but when the estimation model was correctly specified, this power term was 0.32. In fact, the over-specified estimation models, on average, had the greatest power to detect a nonzero effect, with a power of 0.38. This suggests that even if a time series trend is not present in the population, including a time series trend may improve power. On the other hand, the estimation model had no impact

on type 1 error rates for testing the intercept and slope terms of a latent growth model.

Following these analyses, Murphy, Beretvas, and Pituch (2011) incorporated autoregressive and ARMA(1,1) processes to generate and estimate second-order latent growth models. Varying the sample size, the length of the time series, and the generation and estimation methods, the effects of model mis-specification on parameter estimation were quantified as the mean relative bias term, computed as

$$\mathrm{RPB}\left(\hat{\theta}\right) = \frac{\hat{\hat{\theta}} - \theta}{\theta} \qquad (22)$$

where θ is the true parameter value and $\hat{\hat{\theta}}$ is the mean parameter estimate. According to Hoogland and Boomsma's (1998) work, a mean relative bias term greater than 0.05 in absolute value can be considered substantially biased.

When the data were generated using an autoregressive process similar to Equation 14, the mean relative bias of the slope and intercept terms in the latent growth models had less than 0.05 mean relative bias. However, the data generation method had an impact on the estimation of the variance of the slope and intercept terms. That is, when a large autocorrelation was used to generate the data (with $\underline{\varphi}$ = 0.8), estimating the data without an autoregressive term resulted in severe underestimation (of degree 92.1% up to 94.1%) of the variance of the intercept. On the other hand, correctly estimating the data with an autoregressive term resulted in mean relative biases between -0.332 and -0.719. Although handling autoregressive data using a model that correctly specified the autoregressive

structure did improve the mean relative bias of the variance of the intercept term, these estimates were still regarded as being substantially biased, according to Hoogland and Boomsma's (1998) work.

Estimating the variance of the slope was also substantially biased when a large autoregressive term was used to generate the data. Again, the bias was reduced when the estimation correctly included an autoregressive term. Specifically, when the estimation did not include an autoregressive term, the mean relative bias of the slope variance was between -0.616 and -0.703, whereas when the estimation did include an autoregressive term, mean relative bias was between -0.044 and -0.300.

Finally, the estimated covariance between the intercept and slope was substantially biased when an autoregressive structure was simulated, regardless of the size of the generating autoregressive term or the estimation method. When the model was estimated without an autoregressive term, mean relative bias ranged between -0.155 and -0.993, depending mostly on the size of the autoregressive term. However, when an autoregressive estimation model was estimated, the absolute mean relative bias ranged between 0.027 and 1.972, depending mostly on the sample size of the observations.

To summarize, estimating an autoregressive term when an autoregressive term was present resulted in somewhat less biased variance terms for the intercept and slope in Murphy et al.'s (2011) work. However, even when the autoregressive structure was correctly specified, estimates of these terms were substantially

biased. Moreover, when the generating autocorrelation was low ($\varphi = 0.3$), estimation of between 14.3% and 43.7% of the models which included an autoregressive parameter failed to converge, depending on the sample size. Also, when the generating autocorrelation was high ($\varphi = 0.8$), estimation of between 1.3 and 12.7% of the autoregressive models failed to converge, again depending on the sample size.

This proportion of non-convergent cases, along with the substantially biased variance terms, should lead researchers to consider alternative estimation methods. Murphy et al. (2011) used MPlus to estimate these second-order autoregressive latent growth models, which employs likelihood-based estimation, relying on iteratively updating proposed model estimates to check a discrepancy function (Muthén, & Muthén, 2007). These discrepancy terms may achieve a local minimum; furthermore, there is no theoretical guarantee that the estimation method will result in a converged solution.

In contrast, Bayesian methods, and in particular the Gibbs sampler, are theoretically guaranteed to converge eventually (Gelman, Carlin, Stern, & Rubin, 2014). In Gibbs sampling, the conditional distribution of each parameter given all the other parameters is obtained. Then, parameters are initialized at particular values. Next, parameters are sampled according to the conditional distribution, plugging in the most recent sample values for each of the other parameters. This process repetitively continues until, eventually, distributions for each parameter are obtained. Often, the procedure starts with some number of independent samples of values, called chains, and samples are taken from each of the chains separately

(Brooks & Gelman, 1998). Then, the variance in the samples between and within chains is compared to estimate a statistic called the scale reduction factor which provides evidence that the chains have converged (Brooks & Gelman, 1998). Once the chains have converged, one may obtain point estimates of parameters, for instance from the median value of the simulated posterior distribution, or obtain credible intervals, which are somewhat similar to confidence intervals in frequentist statistics (Gelman et al., 2014).

Gibbs samplers for estimating models similar to a first-order latent growth model already exist in the educational psychology literature. For instance, Fox and Glas (2001) outline a Gibbs sampler for a multilevel IRT model. In that work, the researchers were contextualizing the estimation process as a means of estimating an item response theory model for students nested within schools. However, one might use a quite similar algorithm to estimate an item response theory model for measurement occasions nested within examinees. Then, by including a growth term and an intercept for each student, one might apply Fox and Glas's (2001) work to estimate a model similar to Figure 3. However, Fox and Glas (2001) do not explicitly attempt to estimate a model that has an autoregressive component, in fact, it is assumed that observations do not correlate with each other, given the model. As previously discussed, if an autoregressive trend is present in the population and it remains unmodeled, simulation studies indicate that power may be reduced and the actual type 1 error may be inflated for testing growth terms. To account for the

autocovariance structure implied by an autocorrelation term, the Kalman filter is introduced.

Bayesian Methods: The Kalman Filter

Kalman (1960) introduced a filter for sampling latent variables that are autocorrelated. Conceptually, the Kalman Filter proceeds in four steps, the first three of which may be considered together as the forward filter. In the forward filter, first the prior distribution for the latent variable θ_t is obtained, given the previous value of that latent variable, θ_{t-1}. Next, the likelihood for the observed data y_t is obtained, once again given θ_{t-1}. Finally, with the prior and the likelihood for θ_t, the posterior distribution for θ_t is obtained, given the data y_t. The forward filter proceeds in this manner until posterior distributions are obtained for latent variables θ_1, θ_2, …, θ_T. The final step of the Kalman filter, backwards sampling, updates the value for each latent variable θ_t given the next value of that latent variable, θ_{t+1}. The exact mathematics of the Kalman filter will now be delineated; similar derivations may be found in Hamilton (1994), or also in Petris, Petrone, and Campagnoli, (2009).

Consider the following time series structure:

$$y_t = \theta_t + \varepsilon_t \qquad (23)$$

$$\theta_t = \alpha + \varphi * \theta_{t-1} + \zeta_t \qquad (24)$$

with observed variable y_t, latent variable θ_t, error vectors ε_t and ζ_t, latent intercept α, and autoregressive parameter φ. Note that this is a particular application of

Equations 12 and 13. For the sake of stationarity, it is assumed that φ is less than one in absolute value, and it is assumed that the error terms ε_t and ζ_t are uncorrelated at each time point. Also, it is assumed that ε_t and ζ_t are Normally distributed, with means of zero and variances of σ^2 and τ^2, respectively. Finally, let D_t be the set containing elements $y_1, y_2, \ldots, y_t$.

In a Gibbs sampler, the conditional posterior distribution of θ_t is sought, given all the other parameters. Thus, it is assumed that α, φ, σ^2, and τ^2 are all known. As mentioned above, first the prior for θ_t is sought, given the posterior distribution of θ_{t-1}. Suppose for the sake of mathematical induction that the posterior for θ_{t-1} given D_{t-1}, called $p(\theta_{t-1} \vee D_{t-1})$, is Normally distributed, with a mean of m_{t-1} and a variance of C_{t-1}. Then, the prior for θ_t given D_{t-1} is given by

$$p(\theta_t \vee D_{t-1}) = \int p(\theta_t \vee \theta_{t-1}, D_{t-1}) * p(\theta_{t-1} \vee D_{t-1}) d\theta_{t-1} \quad (25)$$

for the previous data D_{t-1}, previous latent estimate θ_{t-1}, and current latent estimate θ_t. Now, by assumption, the posterior distribution $p(\theta_{t-1} \vee D_{t-1})$ is Normally distributed, with a mean of m_{t-1} and a variance of C_{t-1}. Note also that $p(\theta_t \vee \theta_{t-1}, D_{t-1})$ is given by Equation 24, that is, θ_t is Normally distributed, with a mean of $\alpha+\varphi*\theta_{t-1}$ and a variance of τ^2. This integration, then, is analogous to computing the posterior distribution of a variable with a Normal prior, given by $p(\theta_{t-1} \vee D_{t-1})$, and a Normal likelihood, given by $p(\theta_t \vee \theta_{t-1}, D_{t-1})$. A common result in Bayesian statistics (e.g., Gelman, et al., 2014) is that, after computing this integration, $p(\theta_t \vee D_{t-1})$ is Normally distributed, with a mean of $\alpha+\varphi m_{t-1}$ and a variance equal to the quantity $\varphi^2 C_{t-1}+\tau^2$; call these quantities a_t and R_t, respectively.

In the next step of the Kalman filter, the predictive distribution for y_t, given D_{t-1} is obtained in a similar fashion:

$$p(y_t \vee D_{t-1}) = \int p(y_t \vee \theta_t, D_{t-1}) * p(\theta_t \vee D_{t-1}) d\theta_t \quad (26)$$

with $p(y_t \vee D_{t-1})$ the predictive distribution for y_t given the previous data D_{t-1}. Note that the second term of the integration, $p(\theta_t \vee D_{t-1})$, was just shown to be a Normal distribution, with a mean of a_t and a variance of R_t. Next, recall that $p(y_t \vee \theta_t, D_{t-1})$ is given by Equation 23, where y_t is Normally distributed, with a mean value equal to the expected value of θ_t, given D_{t-1}, and a variance equal to σ^2. Above, it was shown that the expected value of θ_t is equal to m_t. Thus, the term in the integration is equal to two Normal distributions, each with the same mean, one with a variance of σ^2 and one with a variance of R_t. The result of such an integration (shown, e.g., in Gelman et al., 2014) is a Normal distribution, with a mean equal to a_t and a variance equal to σ^2+R_t; call this variance term Q_t.

To obtain the posterior distribution for θ_t given D_t, consider the bivariate normal distribution of θ_t and y_t, given D_{t-1}. Once this bivariate Normal distribution is obtained, it will be possible to express θ_t given D_t. Since the univariate Normal distributions of θ_t and y_t given D_{t-1} were obtained above, it only remains to compute the covariance of y_t and θ_t, given D_{t-1}. Recall that this covariance term is defined as

$$\mathrm{Cov}(y_t, \theta_t) = E[(y_t - E\{y_t\}) * (\theta_t - E\{\theta_t\})] \quad (27)$$

where $E(x)$ denotes the expected value of the variable x. Now, $E(y_t) = E(\theta_t)$, as described above, and, by Equation 23, $y_t = \theta_t + \varepsilon_t$. In that case, Equation 27 may be rewritten as

$$\mathrm{Cov}(y_t, \theta_t) = E[(\theta_t + \varepsilon_t - E\{\theta_t\}) * (\theta_t - E\{\theta_t\})] =$$

$$\mathrm{Var}\{\theta_t\} + E[\varepsilon_t * (\theta_t - E\{\theta_t\})] \qquad (28)$$

where the second term may be rewritten as

$$E[\varepsilon_t * (\theta_t - E\{\theta_t\})] = E[\varepsilon_t * \xi_t] \qquad (29)$$

and, by assumption, ε_t and ξ_t are independent, meaning that the expectation of their product is equal to the product of their expected values, which are both equal to zero. In short, the covariance of y_t and θ_t is equal to the variance of θ_t, which was called R_t. In that case, the bivariate normal distribution of y_t and θ_t given D_{t-1} is a bivariate Normal distribution with mean vector $[a_t, a_t]^T$ and variance-covariance matrix $\begin{bmatrix} R_t & R_t \\ R_t & Q_t \end{bmatrix}$. Following standard results on the bivariate Normal distribution, (see. e.g., Jensen, 2000), the posterior distribution of θ_t given D_t is itself Normally distributed, with a mean m_t equal to $a_t + \frac{R_t}{Q_t}$ and a variance C_t equal to $R_t - R_t * \frac{R_t}{Q_t}$. In short, if the Kalman filter is initialized with a variable θ_0 with a particular fixed mean m_0 and a particular variance C_0, the forward filter is able to obtain closed-form solutions for the conditional distribution of each latent variable θ_t.

However, in an autoregressive structure, a latent variable θ_t also depends on the next data point y_{t+1}. In the backwards sampling step of the Kalman Filter, this

dependency is exploited and posterior distributions for θ_t are updated based on y_{t+1}. To see that the value θ_t depends on y_{t+1}, consider the bivariate Normal distribution of θ_t and y_{t+1} given D_t. As discussed above, the mean and variance of θ_t given D_t is m_t and C_t, respectively. Also, the mean and variance of y_{t+1} given D_t is a_{t+1} and Q_{t+1}, since the distribution of θ_{t+1} given D_t is exactly the same as the that distribution sought in the second step of the Kalman filter, above. To fully specify this bivariate Normal distribution, consider the covariance between θ_t and y_{t+1}. By the properties of covariance, and by Equation 24,

$$\text{Cov}(y_{t+1}, \theta_t) = \text{Cov}(\alpha + \varphi\theta_t + \zeta_{t+1} + \varepsilon_t, \theta_t)$$

$$\text{Cov}(\varphi\theta_t, \theta_t) = \varphi\text{Var}(\theta_t) \qquad (30)$$

with the next-to-last equality following from the assumption that the error terms at each time point are uncorrelated with each other. Then, the bivariate Normal distribution of θ_t and y_{t+1} has a mean vector of $[m_t, a_{t+1}]$ and a variance-covariance matrix equal to $\begin{bmatrix} C_t & \phi C_t \\ \phi C_t & Q_t \end{bmatrix}$. Again using standard results about the bivariate Normal distribution, the mean of θ_t given D_{t+1} is given by $m_t + \varphi\frac{C_t}{Q_t}(y_t - a_t)$ and its variance is equal to $C_t - \varphi^2\frac{C_t^2}{Q_t}$. Recall that without this backward sampling step, the variance was equal to C_t, meaning that the subtracted term is variance in the latent estimate that is accounted for by this backward sampling step.

Multivariate extensions of the Kalman filter, along with adaptations for also including moving average processes are available in the literature (e.g., Hamilton,

1994). Taken together, the Kalman filter and the work by Fox and Glas (2001) may be used to construct a Gibbs sampler for a second-order autoregressive latent growth model. However, the work by Fox and Glas (2001) focuses mainly on the recovery of item parameters, while the parameters of interest in latent growth modeling are the latent intercept and growth terms. As a result of these different research aims, a slightly different estimation method, and in particular the use of informative priors, is proposed for the recovery of item parameters.

Informative Priors for Items

In Fox and Glas (2001), item responses are binary, indicating, for example, whether or not an examinee correctly answered a question. Recall Equation 9, reprinted here for convenience:

$$P(y=1) = \Phi(\gamma_1 + \gamma_2 * \eta_2) \qquad (31)$$

where all notation follows Equation 9. Following Fox and Glas (2001), γ_1 is analogous to the vector of difficulty parameters, b, in item response theory, whereas γ_2 is analogous to the vector of discrimination parameters, a, in item response theory. Finally, the vector η_2 represents what is commonly referred to in item response theory as the vector θ of ability estimates.

In the original work (Fox and Glas, 2001) the prior distributions for a and b were non-informative priors, specifying only that every value of a was greater than zero. More explicitly, the priors for a and b, as written in Fox and Glas (2001) are given as

$$p(a)p(b) \propto \prod_{k=1}^{K} I(a_k > 0) \qquad (32)$$

where the function in the product term is equal to one if a_k is greater than zero and equal to zero otherwise. This prior term represents that, conventionally in item response theory, the *a* terms are typically greater than zero; no other information is contained in Equation 32.

However, if the modeling of latent growth is of primary concern, and not necessarily the calibration of each item, it may be possible to use informative priors instead. Supposing that the items were calibrated on an independent sample, researchers might use those calibration statistics in informative priors. These calibration statistics may be obtained, for instance, using marginal maximum likelihood estimation (Rizopoulos, 2006). Given these calibration statistics, researchers have suggested, from mathematical theory and simulation studies, that the use of weakly informative priors may lead to a reduction in mean squared error for parameter estimation, (e.g., Yuan & MacKinnon, 2009; Gelman, et al., 2014; Kruschke, 2011). Particularly, when 100 or fewer cases are observed, simulation studies have found that weakly informative priors reduce the root mean square error of approximation by up to half (Yuan & MacKinnon, 2009). For parameters with normally distributed priors and likelihoods, the use of informative priors is analogous to adding data points from previous research to the likelihood function of the current dataset (Gelman, et al., 2014). In so doing, one assumes that the data from previous and current research are exchangeable. In general, the plausibility of this assumption depends on the nature of the research involved.

Formally, exchangeability is defined as follows. A set of random variables X_1, X_2, …, X_n are deemed exchangeable if their joint probability distribution does not depend on their subscripts (de Finnetti, 1974; Lindley & Novick, 1981; Rubin, 1978). In other words, the order in which the variables are observed does not impact their joint probability distribution. If using an informative prior is analogous to including data points from previous research, then their use assumes that the data generation mechanism is identical between studies.

For most research in the social sciences, this may be a strong assumption. However, for item-response theory parameters, the assumption is justifiable. It has long been thought that item response theory makes item calibration person-free; that is, given a large enough sample size for item calibration and model fit, item calibration does not depend on the ability level of the persons in the calibration sample (Wright, 1967). Thus, item parameter data from the calibration sample should be exchangeable with those parameter data from the sample being analyzed. In short, exchangeability is not a particularly strong assumption within the context of item response theory, and so informative parameters may be used for item parameters.

Study Two: Empirical Data Analysis

For illustrative purposes, a second-order autoregressive latent growth model using a recently collected dataset was also estimated. Particularly, the study obtained data on depression and suicidal ideation from gay, lesbian, bisexual, trans, and questioning (LGBTQ) youth between 15 and 20 years of age collected in four

separate waves across three cities; more details on the sample may be found in the original study (Baams, Grossman, and Russell, 2015). The data include a variety of validated measurements, including, among others, the Interpersonal Needs Questionnaire, or INQ (Van Orden, Witte, Gordon, Bender, and Joiner, 2008).

The INQ contains two major subscales, namely, perceived burdensomeness and thwarted belongingness, both of which are empirically associated with suicidal ideation (Van Orden et al., 2008). Perceived burdensomeness is defined as the extent to which youth feel that they are a burden on family and friends, while thwarted belongingness refers to an unmet need to belong. While the original study (Van Orden et al., 2008), exclusively evaluated these factors as distinct constructs, the more recent research (Baams et al., 2015) found that the two scales correlated at 0.53, indicating a moderate-to-strong relationship between the two constructs. The validation study (Van Orden et al., 2008) found strong internal consistency estimates for thwarted belongingness (alpha = 0.85) and perceived burdensomeness (alpha = 0.89).

Recent empirical research (Baams et al., 2015) found evidence to suggest that perceived burdensomeness, in particular, may be an important mediator related to depression and suicidal ideation. Given its empirical relevance, perceived burdensomeness was selected for a real data analysis using a second-order autoregressive latent growth model. The stability of this construct, as quantified by its autoregression, may be relevant to researchers. For instance, perceived burdensomeness could be thought of as a marker for short-term suicide risk if the

autoregression is relatively low, or a longer-term marker if the autoregression is quite high. Either way, the stability of the construct, and the extent to which this stability varies across people, may be of empirical interest.

Statement of Purpose

Murphy et al. (2011) investigated the estimation of a second-order autoregressive latent growth model, with results indicating generally underestimated variance terms and high rates of non-convergence even when the correct model was estimated. The proposed estimation method may improve upon this analytic method in two distinct ways. First, the proposed method is Bayesian in nature, and is therefore theoretically guaranteed to converge eventually. The method incorporates the Kalman filter, a method specialized for time series data. Secondly, the proposed analysis uses previously calibrated item parameters as informative priors. Both mathematical theory and simulation studies suggest that this feature should result in a reduction in mean squared error for parameter estimates. Furthermore, the use of fully informative priors is theoretically justifiable, since item calibration is person-free (Wright, 1967).

While similar to previous work (Fox & Glas, 2001), the proposed model differs from Fox and Glas's (2001) analysis in two ways. Recall that Fox and Glas (2001) proposed a multilevel IRT model using Gibbs sampling. The proposed model uses measurement occasions at the first level of the multilevel model, and person-wide ability estimates at the second level. However, unlike Fox and Glas's (2001) analysis, this work includes an autoregressive component, which converges more efficiently using the forward filtering backwards sampling algorithm. Secondly, unlike Fox and Glas (2001), the present analysis focuses more on the analysis of

change over time, and less on the calibration of items. As such, informative priors are employed for item parameter estimates.

After the Gibbs sampler is developed, a simulation study will be presented to assess the utility of these modifications. The sample size for the calibration of the items are varied, along with the sample size for the latent growth analysis. Moreover, the priors for the item parameters are varied, so that priors are strongly informative, weakly informative, or non-informative. Mean squared error and mean relative bias for person-level ability estimates, item response theory parameters, and the means and variances of the growth and the autoregressive trend are assessed across conditions. The 95% credible interval coverage are assessed for these parameters also. Based on mathematical theory, it was hypothesized that RMSEs, credible interval coverage rates, and average bias would decrease for person-level ability estimates as the item calibration sample size increases, as the latent growth analysis sample size increases, and as priors for ability estimates become more informative.

There are five research questions about second-order autoregressive latent growth models that the first study aims to address:

1. To what extent do greater sample size, lower autoregression, and more informative priors for item parameters improve convergence?
2. To what extent do greater sample size, lower autoregression and more informative priors for item parameters reduce the mean relative bias of the latent growth's mean and variance?

3. To what extent do more informative priors for item parameters reduce the mean squared error of estimation of the latent growth's mean and variance?
4. To what extent do more informative priors for item parameters improve credible interval coverage rates of the latent growth's mean and variance?
5. To what extent do greater sample size, smaller autoregression, and more informative priors for item parameters improve overall parameter recovery?

After the results for the simulation study were obtained, a brief empirical analysis estimating perceived burdensomeness in LGBTQ youth was conducted. The main purpose of this empirical analysis was to briefly identify practical problems that may arise in the analysis of real data. However, since perceived burdensomeness was found to be an important mediator predicting suicidal ideation, (Baams et al., 2015), the empirical analysis also addressed an applied research question. Mainly, to what extent is perceived burdensomeness stable over time?

Methods for study one

Data Generation: Latent Growth Model Sample

A simulation study was conducted to assess the bias and mean squared error of parameters estimated using the proposed estimation method for a second-order autoregressive Bayesian latent growth model.

Most of the generating parameters in this analysis are specified in Equation 18, reprinted here for convenience:

$$\begin{bmatrix} 1 \\ t+1 \\ \theta_{t+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ \alpha & \beta & \varphi \end{bmatrix} \begin{bmatrix} 1 \\ t \\ \theta_t \end{bmatrix} + \zeta_t \qquad (33)$$

with the time parameter t coded as zero for the first measurement occasion, one for the second measurement occasion, and so on. Persons' ability levels at the first time point, or a in the above, will be generated in two steps. First, person-level ability estimates were taken as random draws from the standard normal distribution. This process was intended to simulate differences in initial ability. For identification purposes, the estimation psrocedure set initial person-level ability as a standard Normal distribution also. Next, occasion-specific error for the first time point was simulated by setting a equal to a draw from the Normal distribution, with a mean at the person-level ability estimate, and variance equal to one-third, corresponding to a medium level of occasion-specific variance (from Geiser, Keller, & Lockhart, 2013). That is, the total variance for the first time point was equal to four-thirds,

with a variance of one associated with true variance in ability and a variance of one-third associated with time-specific measurement error.

Next, the autoregressive parameter (φ) in the above was taken as a draw from a truncated normal distribution, with a mean of 0.65 and a variance of 0.273; this approximated the range between 0.5 and 0.8 which was investigated in Price, (2012) for a model analogous to a first-order autoregressive latent growth model.

The growth parameter (β) in the above equation was generated as a draw from a normal distribution, with a mean of 0.34, and a variance of 0.17, corresponding to Rosenbaum's (1986) estimate of high school student growth in mathematics, per semester. Using these values of α, β, φ, and θ_1, future time-specific ability estimations were generated following Equation 18. Matching previous research in second-order latent growth modeling, four time points were simulated (Geiser, Keller, & Lockhart, 2013).

Finally, once values for θ_t were generated, individual items for specific persons at specific time points were generated using the following formula:

$$P(\mathbf{y}=1) = \Phi(a^*\theta_t - b) \qquad (34)$$

with $\Phi(x)$ representing the cumulative density function of the standard normal distribution, b representing the vector of difficulty parameters, and a representing the vector of item discrimination parameters. Values for the 30 items' discrimination and difficulty parameters will match those reported for a previously calibrated ACT mathematics test and were used to generate binary responses at

each time point. In other words, the item parameters were simulated as being constant over time, with no item drift or any substitutions of different items at different times. These item parameters are reported in Table 1, below.

Table 1. Item parameters used in the simulation study.

	Item Parameter	
Item number	Discrimination *A*	Difficulty *b*
1	0.684969	-1.97243
2	0.847535	-1.00585
3	0.701955	-1.22521
4	0.902016	-1.36332
5	1.046108	-0.28199
6	0.890067	-0.57313
7	0.893602	-1.36773
8	0.716244	-0.85472
9	1.164358	-0.44410
10	0.368682	-0.32641
11	0.591835	0.074493
12	0.965367	0.030348
13	0.829711	0.181143
14	1.442186	0.212373
15	1.039885	-0.28632
16	0.905395	-0.04697
17	1.457840	-0.01277
18	1.222484	0.171493
19	0.867419	0.204116
20	1.006531	0.324807
21	0.948628	0.345869
22	0.855386	1.452706
23	0.999414	0.731968
24	1.136003	1.229729
25	0.804661	1.273888
26	0.967754	1.803751
27	1.240903	1.047733
28	1.589016	1.626541
29	1.446432	1.895972
30	1.364186	2.806588

Data Generation and Estimation: Item Calibration

Simulating initial item calibration proceeded in three steps. First, ability levels θ_I were generated from a Standard Normal distribution. Then, responses were generated following Equation 34, with values for a and b equal to those used in generating the Latent Growth Model simulated dataset. In other words, it was

assumed that items are calibrated at one specific time point, and that the item parameters do not change over time. Finally, item parameters were estimated using the package "ltm" in R (Rizopoulos, 2006). Depending on the condition, these calibration statistics may be used in informative priors for item parameters.

Simulation Study Conditions

The conditions consisted of a three by three by three factorial design, with a fully balanced combination of the following three factors. First, three sample sizes were varied for the latent growth model, with simulee sample sizes of 100, 500, and 1,000 (the latter matching sample sizes used in Geiser, Keller, & Lockhart, 2013). The smaller sample sizes are included to evaluate the effects of informative priors when used with smaller sample sizes. Also, the next condition will vary the mean value of φ, with mean values of 0, 0.3, and 0.8. Finally, non-informative, weakly informative, and strongly informative priors will be used to estimate the model.

Non-informative priors are specified identically to those used in Fox and Glas's (2001) work. In the strongly informative prior condition, the calibration statistics from the subsection above were used as the informative priors for item parameters. That is, the prior for each item parameter was taken as normally distributed, with the mean and variance based directly on the values of the point and the squared standard error estimates calculated in the calibration. Finally, in the weakly informative prior condition, the variance of the item parameters was taken as three times the squared standard error given from the calibration.

Outline of the Gibbs Sampler

To run this Gibbs Sampler, the software packages R and Rcpp were used. As previously stated, the Gibbs Sampler was largely taken from Fox and Glas's (2001) work, with only two major adjustments. Mainly, rather than the first level corresponding to students and the second level corresponding to schools, in this analysis the first level corresponds to measurement occasions and the second level corresponds to examinees. The measurement model is given by

$$P(\mathbf{y}=1) = \Phi(b - a*\theta_t) \qquad (35)$$

with notation identical to Equation 34. Then, the first level of the model is given by

$$\begin{bmatrix} 1 \\ t+1 \\ \theta_{t+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ \alpha & \beta & \varphi \end{bmatrix} \begin{bmatrix} 1 \\ t \\ \theta_t \end{bmatrix} + \zeta_t \qquad (36)$$

with notation the same as Equation 33. Call the last element of ζ_t e_t; then, e_t is assumed to be Normally distributed, with a mean of zero and a variance called σ^2. Finally, in the second level, the terms α, β, and φ are assumed to follow a multivariate Normal distribution, with means of $\lambda_{\alpha 0}$, $\lambda_{\beta 0}$, and $\lambda_{\varphi 0}$, respectively, and covariance matrix $\mathbf{T}$; call the vector containing the terms α, β, and φ, β_i for a given individual i.

Before detailing the algorithm discussed in Fox and Glas (2001), the adjustments made to the algorithm are mentioned first. The adjustments are threefold. First, for those conditions with informative or weakly informative priors, the item parameters will have normal priors. Following Equation 12 in Fox and

Glas (2001), the item parameters have normal likelihood functions also. Without loss of generality, consider a particular item discrimination parameter a. Say that, based on the item calibration, the prior for a is specified as normally distributed, with mean a_p and standard error σ_p. Also, call the ordinary least squares estimator for this parameter $\hat{a}$, with standard error $\hat{\sigma}$. Then, for each item, the fully conditional posterior density for a is given by

$$\alpha \sim N\left(\frac{\hat{a}/\hat{\sigma}^2+a_p/\sigma_p^2}{1/\hat{\sigma}^2+1/\sigma_p^2}, \frac{1}{1/\hat{\sigma}^2+1/\sigma_p^2}\right) \qquad (37)$$

with $\sim N$ specifying that the term is normally distributed, with the given mean and variance. Item difficulty terms b have exactly analogous conditional posterior density functions.

Secondly, the latent constructs θ are sampled using the filter-forward backward sampling algorithm, that is, the Kalman filter as described in the literature review above. Thirdly, the autoregressive parameter φ are estimated separately from the other level-two regression weights, since, for stationarity purposes, φ must be constrained to be less than one in absolute value. To accomplish this, φ were estimated assuming knowledge of every other parameter and sampled as a truncated Normal distribution, with truncation points at negative and positive one. The mean and variance of this truncated Normal distribution are computed identically to how these terms are computed for the other level-two regression weights, described below.

Following Fox and Glas, (2001), in the first step latent variables Z are sampled using truncated Normal distributions, each with means of $b + a^* \theta_t$ and variances of one. To reflect the binary nature of the response data, the latent variables Z are constrained to be greater than zero if the corresponding question was correctly answered and are constrained to be less than zero otherwise. This is a probit link to the binary data.

In the next step, estimates θ_t were sampled using the Kalman filter, as described above. For more details, see the section on this topic in the literature review, above.

In the third step, the item parameters in the a and b vectors were sampled, using Equation 35, above, depending on condition. For those conditions in which informative priors are not provided, these parameters are taken as a draw from the multivariate Normal distribution. Let $\boldsymbol{\xi_k}$ be the vector containing item parameters a_k and b_k, respectively, and let $\mathbf{Z_k}$ be the vector containing all those elements of Z corresponding to item number k. Then, as per Fox and Glas (2001),

$$\mathbf{Z_k} = [\ \boldsymbol{\theta} \ \mathbf{-1}\ ]\ \boldsymbol{\xi_k} + \boldsymbol{\varepsilon_k} \qquad (38)$$

where $\boldsymbol{\varepsilon_k}$ is a Normally distributed random variable with a mean of zero and a variance of one. The vector $\boldsymbol{\xi_k}$ may then be sampled as a multivariate Normal, with mean $\hat{\xi}_k$ corresponding to Ordinary Least Squares regression weights and covariance matrix $([\ \boldsymbol{\theta} \ \mathbf{-1}\ ]^{T} * [\ \boldsymbol{\theta} \ \mathbf{-1}\ ])^{-1}$.

After the item parameters are sampled, the terms α, and β are sampled, followed by separately sampling φ, as described above. In the former step, α and β

were sampled again according to Fox and Glas (2001). Define $\mathbf{d} = \sigma^2 * \boldsymbol{\theta} + \mathbf{T_a}^{-1} \gamma$, where $\boldsymbol{\gamma}$ is the vector containing γ_α and γ_β, and $\mathbf{T_a}$ is the principal submatrix of $\mathbf{T}$ formed by omitting the final row and column. Next define $\mathbf{D} = (1/\sigma^2 + \mathbf{T_a}^{-1})^{-1}$. Then, the terms α and ß are sampled from a multivariate Normal distribution with mean $\mathbf{Dd}$ and covariance matrix $\mathbf{D}$.

Next all the elements of $\boldsymbol{\gamma}$ were sampled, continuing to mirror the algorithm described in Fox and Glas (2001). In this application, define $\mathbf{E} = \mathbf{T} * 1/n$, and define $\mathbf{e} = \sum_{i=1}^{N} T^{-1} * \beta_i$. Then, $\boldsymbol{\gamma}$ is drawn as a multivariate Normal distribution, with mean $\mathbf{Ee}$ and covariance matrix $\mathbf{E}$.

In the second-to last step, the error variance term σ^2 is sampled, again following Fox and Glas (2001). Let $\mathbf{X}_t$ equal [1, t, θ_t], for a particular time point t. Then, let S^2 equal $\frac{1}{n}\sum_{t=1}^{T}(\theta_t - X_t\beta)^T(\theta_t - X_t\beta)$. The error variance is then taken as an inverse-gamma distribution, with shape parameter N and scale parameter S^2.

Finally, the level-two covariance matrix is sampled in a roughly similar manner to the previous step. Fox and Glas (2001) define a variable $\mathbf{S}$ equal to $\sum_{t=1}^{T}(\beta - \gamma)^T(\beta - \gamma)$. Then, the covariance matrix $\mathbf{T}$ is sampled as an inverse-Wishart distribution, with t degrees of freedom and scale matrix $\mathbf{S}^{-1}$.

All of these steps together constitute a single iteration of the Gibbs sampler. As described below, these iterations were repeated some 11,000 times for each time the model is estimated. Finally, three distinct chains were run for each model to allow for convergence checking using the Gelman-Rubin statistic (Gelman & Rubin,

1994). A model was considered converged only when the Gelman-Rubin statistic is less than 1.1 for all parameters; the number of nonconverged cases were tallied. To further assess convergence, the traceplots of each parameter were evaluated once per condition. That is, for each condition, one initial estimation was obtained, and the traceplots for each parameter was evaluated.

Pre-testing

Since this estimation procedure used new software in R, many pre-tests were conducted before simulation proceeded. Essentially, particular components of the model were systematically generated and estimated in isolation. Obviously incorrect estimates necessitated debugging and further testing. When the estimation of several specific parameters were verified in this way, those parameters were then estimated together in further tests.

For instance, in one pre-test, θ_t values were generated according to Equation 33, above, with 1000 simulees and 4 time points. In that pre-test, the parameters α, β, and φ were treated as known. In other words, only the Kalman filter was used in the estimation process, checking the estimation of the θ_t values. In the next pre-test, the θ_t values were treated as known so that the α, β, and φ estimates could be estimated. After both portions of code were checked in this way, the next pre-test involved estimating θ_t, α, β, and φ simultaneously.

Estimation of every parameter in the model was tested in this fashion, gradually adding new parameters to a larger estimation model after they were

separately pre-tested. Each pre-testing used 1000 simulees, 4 time points, a true φ value of 0, and non-informative priors for item parameters. In the final stages of pre-testing, three large portions of the estimation were checked. Particularly, the IRT portion of the model in Equation 34 was checked, simulating and estimating the a, b, θ, and latent Z variables, treating every other parameter as known. Secondly, the Kalman filter was tested, obtaining θ values with known Z, α, β, and φ values. Finally, the hierarchical linear portion of the model was tested, estimating α, β, φ, $\boldsymbol{\gamma}$, and $\mathbf{T}$ with known θ values. Although results for these pre-tests are not reported for the sake of brevity, the performance of each of these portions of the estimation was thought to be adequate to continue.

Simulations and Analysis

Following Geiser, Keller, and Lockhart (2013), for each condition, 1,000 datasets were simulated. For each dataset, following Price (2012), there were 1,000 burn-in iterations, and 10,000 samples from the posterior distribution. Once the posterior distributions were obtained, the mean squared error and mean relative bias for person ability, growth, autoregressive components, item response theory parameters, level-one variance, and the level-two variances of growth and the autoregressive trend were computed, with the point estimate taken as the median of the posterior distribution. After these statistics were obtained, the mean squared error statistics were used to compare parameter recovery across the uninformative, partially informative, and fully informative priors for IRT parameters, and across the different generating sample sizes and φ values. It was hypothesized that the

more informative priors would improve parameter recovery by reducing the mean squared error. Also, the mean relative bias estimates assessed parameter recovery across all conditions. I hypothesized that greater sample sizes and more informative priors would result in lower mean relative bias, overall. Also, coverage rates for the 95% credible intervals were computed for these same parameters. Again, it was hypothesized that coverage rates would improve with greater sample sizes and more informative priors.

Results for study one

Computing time

Although not directly related to the primary research questions, due to the extreme computational difficulties involved with the simulation study, its computing time is reported here. The simulation study was conducted using two separate supercomputers, stampede and stampede2, both provided by the Texas Advanced Computing Center (TACC, 2017). Stampede was used from June 1st, 2016, to June 6th, 2017, while stampede2 was used from June 7th, 2017, to November 15th, 2017. During this time, a total of 409,382.45 service units (SUs) were used on stampede, and 10,554.10 SUs were used on stampede2.

Service units were defined almost identically for stampede and stampede2. Full specifications for these supercomputers are available online (TACC, 2017), however the SUs are briefly defined here. On stampede, the SUs were defined as the number of computing nodes used multiplied by the number of cores per node multiplied by the total duration of the job. On stampede2, the SUs were defined similarly, except the number of cores per node was not a multiplier. At the time, every computing node on stampede contained two 8-core processors, for a total of 16 cores, each of which was associated with 2GB of RAM, for a total of 32 GB of RAM per node. Across both stampede and stampede2, jobs were run in parallel using as many cores as the RAM would support. Therefore, the stampede jobs alone would have taken 812,764.9 hours to finish on a computer with 16 GB of RAM, assuming a processor capable of an equal number of floating-point operations per second. This

estimate was obtained by doubling the number of SUs, although naturally the actual number of hours would depend on the precise specifications of the CPU and RAM of this hypothetical computer.

Stampede2 had more cores and more RAM associated with each computing node. On that system, only the phase 1 computing nodes were used, each of which contained 68 cores, associated with a total of 96 GB of RAM. Performing a similar calculation as above, the SUs expended on stampede2 would amount to roughly 63,324.6 computing hours on this hypothetical computer with 16 GB of RAM and an equally powerful CPU.

In total, then, the jobs corresponded to roughly 876,089.5 hours of computing time for a machine with 16 GB of RAM, or just over 100 years of computing time. Due to these difficulties, only 500 replications were conducted for the largest sample size (N=1,000) conditions with 1,000 replications used for all other conditions. Additionally, the remaining results are based only on the cells that converged after 500 or 1,000 replications. Convergence rates are described below.

Convergence

To answer the first research question, convergence rates for the simulation study are presented in Table 2, below. Recall that the first research question asked what the convergence rates are, under a variety of conditions, for this latent growth model. The highest convergence rate (91%) was for the condition with N = 1,000 and $\varphi = 0$, with non-informative priors. The lowest convergence rate (3.20%) was found

for the condition with N =1,000 and φ = 0.8, with weakly informative priors. Generally, a higher φ value was associated with a lower convergence rate.

Convergence rates for the weakly and strongly informative conditions appeared to be roughly comparable, but were generally lower than the non-informative conditions. The only exception was for those cells in which the sample size was 100 and the autoregression parameter was 0.8; in those cases, the non-informative cell converged 17.86% of the time, while the condition with a strongly informative prior converged 20.52% of the time. However, the weakly informative condition converged 17.13% of the time, less than the non-informative condition. Despite this exception, on average, the non-informative conditions converged 7.73% more often than the weakly informative conditions and 7.99% more often than the strongly informative conditions. The maximum difference of 21.31% was between the non-informative and strongly-informative conditions, for those cells with N = 500 and φ = 0. For the weakly informative conditions, the biggest difference from the non-informative conditions was in cells with N = 500 and φ = 0.3, with a discrepancy of 19.34%. Detailed information on model convergence may be found in Table 2, below.

Table 2. Convergence rates by condition.

		Prior		
N	φ	Non-informative	Weakly informative	Strongly informative
100	0	39.29%	35.46%	38.53%
	0.3	26.77%	23.52%	27.18%
	0.8	17.86%	17.13%	20.52%
500	0	87.51%	78.06%	66.20%
	0.3	57.93%	38.59%	42.05%
	0.8	14.54%	5.17%	4.70%
1000	0	91.00%	88.00%	84.00%
	0.3	63.60%	46.60%	46.00%
	0.8	6.80%	3.20%	4.20%

Mean Relative Biases

To answer the second research question and partly address the final research question, mean relative bias terms were calculated and are presented in Table 3, below. Recall that the second research question asked what the mean relative bias for the mean and variance of the latent growth term would be using this estimation procedure. Also, the last research question asked under what conditions parameter recovery would be impacted by using more strongly informative priors. Correlations between true and estimated item parameters are reported in the next section.

Across conditions, the average mean relative bias for the mean of the level-two growth term was -0.321, indicating that this growth term was generally underestimated. The cells with non-informative priors were less biased than average, with an average mean relative bias of -0.270, while the weakly and strongly informative priors had mean relative biases of -0.354 and -0.339, respectively. Growth terms were the least biased for higher values of φ, on average;

cells with φ = 0 had an average mean relative bias of -0.368, while the cells with φ = 0.3 had an average mean relative bias of -0.322. Finally, cells with φ = 0.8 had average mean relative biases of -0.273 for the level-two average growth term. Unusually, the mean relative bias was the smallest for the cells with N=100, with an average mean relative bias of -0.289, compared to -0.344 and -0.330 for N =500 and N =1000, respectively.

The average mean relative bias of the variance of the level-two growth term was also generally underestimated, with a grand mean value of -0.524. This mean relative bias was lower, on average, for less informative prior conditions; the term was equal to -0.485, on average for the non-informative prior conditions, -0.519 for the weakly informative prior conditions, and -0.568 for the strongly informative prior conditions. It was lower than average for the conditions with φ = 0.8, with a value of -0.454; the other φ conditions differed by less than 0.01 from each other. Again, the mean relative bias term was the lowest for the N = 100 cells, with a mean relative bias of -0.478; again, the other two sample size biases differed by less than 0.01 from each other.

The paragraphs above provide results pertinent to the second research question. To partly address the final research question, a brief overview of the remaining columns in Table 3 is provided. Again, the second research question asked what the mean relative bias for the mean and variance of the latent growth term would be, while the final research question asked how parameter recovery would be impacted by using more strongly informative priors Generally, the level-

one regression coefficients appeared to be underestimated. The theta estimates were the least underestimated for the non-informative prior conditions, with an average mean relative bias of -0.323, compared to -0.427 and -0.653 for the weakly and strongly informative prior conditions, respectively. The level-one variance was generally overestimated, with an average mean relative bias of 0.489. The level-two variance of the autoregression term was generally severely underestimated, with an average mean relative bias of -0.900. Finally, the mean relative bias of the average autoregression term was generally underestimated, with an average mean relative bias of -0.432. For more information on mean relative biases across conditions, please refer to Table 3, below.

Table 3. Mean relative biases across conditions.

			Level-one regression terms				Variance terms			Level-two means	
N	φ	Prior distributions	α	β	Φ	θ	σ^2	T_{22}	T_{33}	γ_2	γ_3
100	0	Non-informative	0.517	-0.670	-0.063	-0.840	-0.073	-0.525	-0.915	-0.404	-0.184
		Weakly informative	-0.487	-0.523	-0.916	-0.282	-0.103	-0.212	-0.338	-0.381	-0.144
		Strongly informative	-0.419	-0.412	-1.005	-0.730	0.082	-0.727	-0.896	-0.399	-0.063
	0.3	Non-informative	-0.799	-0.509	-0.936	0.074	0.075	-0.509	-0.936	-0.357	-0.868
		Weakly informative	-0.478	-0.542	-0.577	-0.416	1.148	-0.503	-0.878	-0.245	-0.434
		Strongly informative	-0.440	-0.217	-0.770	-0.350	1.377	-0.534	-0.868	-0.234	-0.367
	0.8	Non-informative	-0.372	-0.535	-0.309	0.148	0.274	-0.524	-0.976	-0.325	-0.540
		Weakly informative	-0.264	-0.261	-0.282	-0.295	1.883	-0.413	-0.976	-0.156	-0.353
		Strongly informative	-0.245	-0.286	-0.182	-0.641	2.216	-0.350	-0.979	-0.098	-0.342
500	0	Non-informative	-1.414	-0.914	-0.969	-0.338	0.594	-0.503	-0.824	-0.316	-0.096
		Weakly informative	-0.310	0.233	-0.739	0.063	-0.120	-0.641	-0.888	-0.394	-0.147
		Strongly informative	-0.664	-0.243	-0.962	-2.199	0.026	-0.642	-0.870	-0.373	-0.119
	0.3	Non-informative	-0.348	-0.462	-0.788	-0.416	1.148	-0.503	-0.878	-0.245	-0.434
		Weakly informative	-0.479	-0.544	-0.956	-0.740	-0.167	-0.602	-0.942	-0.402	-0.895
		Strongly informative	-0.286	-0.513	-1.102	-0.185	-0.013	-0.601	-0.927	-0.387	-0.796
	0.8	Non-informative	-0.343	-0.407	-0.056	-0.295	1.883	-0.413	-0.976	-0.156	-0.353
		Weakly informative	-0.441	-0.518	-0.611	-0.286	-0.312	-0.514	-0.990	-0.408	-0.711
		Strongly informative	-0.510	-0.531	-0.580	-0.685	-0.346	-0.518	-0.993	-0.418	-0.704
1000	0	Non-informative	-0.186	-0.348	-1.312	-0.248	0.730	-0.502	-0.813	-0.295	-0.080
		Weakly informative	-0.433	-0.863	-0.572	-1.675	0.026	-0.642	-0.870	-0.373	-0.119
		Strongly informative	-0.471	-0.542	-0.943	-0.391	0.021	-0.646	-0.872	-0.375	-0.120
	0.3	Non-informative	-0.801	-0.154	-0.473	-0.350	1.377	-0.534	-0.868	-0.234	-0.367
		Weakly informative	-0.486	-0.531	-0.956	-0.715	-0.156	-0.624	-0.926	-0.409	-0.918
		Strongly informative	-0.853	-0.878	-0.969	-0.236	-0.024	-0.608	-0.926	-0.387	-0.793
	0.8	Non-informative	-0.247	-0.378	-0.087	-0.641	2.216	-0.350	-0.979	-0.098	-0.342
		Weakly informative	-0.502	-0.497	-0.623	0.506	-0.346	-0.518	-0.993	-0.417	-0.704
		Strongly informative	-0.448	-0.476	-0.532	-0.460	-0.224	-0.484	-0.992	-0.385	-0.658

Item Response Theory Parameter Recovery

To partly address the final research question, which asked how parameter recovery would be impacted by using more strongly informative priors, the recovery of the IRT parameters is reported. Again, the correlations are reported since there was an unexpected scaling problem. The scaling problem is elaborated upon in the discussion section.

Generally, correlations between the true and estimated a_k parameters were over 0.834, and correlations between the true and estimated b_k parameters were over 0.948. Also, with one exception, the median theta values were correlated greater than 0.93 with the generating values. The exception was in the cell with a generating sample size of 100, an autoregression of zero, and with weakly informative priors, which had a correlation of just 0.685. On average, estimated a_k, b_k, and theta estimates were correlated with the generating values at 0.946, 0.971, and 0.942, respectively. These correlations appeared to differ slightly by condition; the average correlation for the a_k parameters was 0.923 for the non-informative conditions, 0.951 for the weakly informative conditions, and 0.963 for the fully informative conditions. However, the estimated b_k parameters correlated at 0.990 for the non-informative conditions, 0.961 for the weakly informative conditions, and 0.962 for the fully informative conditions. See Table 4, below, for more detail on these correlations.

Table 4. Correlations between estimated and true IRT parameters.

N	φ	Prior distribution	a_k	b_k	θ
100	0	Non-informative	0.8342	0.9758	0.9362
		Weakly informative	0.8964	0.9477	0.6845
		Strongly informative	0.9328	0.9485	0.9501
	0.3	Non-informative	0.8427	0.9764	0.9432
		Weakly informative	0.9462	0.9951	0.9493
		Strongly informative	0.9696	0.9973	0.9506
	0.8	Non-informative	0.8703	0.9774	0.9528
		Weakly informative	0.9565	0.9961	0.9584
		Strongly informative	0.9707	0.9973	0.9594
500	0	Non-informative	0.9464	0.9946	0.9450
		Weakly informative	0.959	0.9514	0.9496
		Strongly informative	0.9652	0.9518	0.9494
	0.3	Non-informative	0.9462	0.9951	0.9493
		Weakly informative	0.9566	0.9512	0.9534
		Strongly informative	0.9631	0.9517	0.9537
	0.8	Non-informative	0.9565	0.9961	0.9584
		Weakly informative	0.9547	0.9512	0.9595
		Strongly informative	0.9607	0.9517	0.9601
1000	0	Non-informative	0.9699	0.9971	0.9472
		Weakly informative	0.9652	0.9518	0.9493
		Strongly informative	0.969	0.9521	0.9492
	0.3	Non-informative	0.9696	0.9973	0.9506
		Weakly informative	0.9641	0.9517	0.9519
		Strongly informative	0.9681	0.9521	0.9527
	0.8	Non-informative	0.9707	0.9973	0.9594
		Weakly informative	0.9607	0.9517	0.9601
		Strongly informative	0.9637	0.9518	0.9607

Mean Squared Error

To answer the third research question and partially address the final research question, the mean squared error of various parameters across conditions are reported. Complete information on these statistics may be found in Table 5, below. Once again, the third research question investigated the mean squared error

for the mean and variance of the level-two growth terms, while the final research question addressed general parameter recovery.

Addressing the third research question, the level-two average growth terms, on average, had a mean squared error of 0.0142. The mean squared error increased slightly as priors became more informative; the mean squared error was 0.0106 for the non-informative conditions, 0.0171 for the weakly informative conditions, and 0.0150 for the strongly informative conditions. The mean squared error decreased as the autoregression factor, φ, increased. With a level-two autoregression of zero, the average mean squared error was 0.0179, whereas with an autoregression of 0.3, the average mean squared error was 0.0134. Finally, with an autoregression of 0.8, the average mean squared error was 0.0114. The mean squared error differed by less than 0.0015 when sample size was varied. Specifically, with an N of 100, the average mean squared error was 0.0136, and with an N of 500, the average mean squared error was 0.0150. With the largest sample size of 1000, the average mean squared error was 0.0141.

The average mean squared error for the variance of the level-two growth term was generally slightly lower, by comparison; the average mean squared error of T_{22} was 0.0093. Again, the mean squared error increased as priors became more informative, with an average mean squared error of 0.0083 for the non-informative conditions, 0.0093 for the weakly informative conditions, and 0.0102 for the strongly informative conditions. Additionally, the average mean squared error decreased as the autoregressive parameter increased. There was an average mean squared error

of 0.01096 for conditions with φ equal to 0, 0.0010 for conditions with φ equal to 0.3, and 0.007 for conditions with φ equal to 0.8. The mean squared errors varied by less than 0.0003 as the sample size changed. With a simulation sample size of 100, the average mean squared error was 0.00928, whereas a simulation sample size of 500 had an average mean squared error of 0.0094. Finally, with the largest sample size of 1000, the average mean squared error was 0.0091.

Partly addressing the final research question, which addressed general parameter recovery, I report the mean squared errors for the person-level parameters, including person-level ability estimates, intercepts, slopes, and autoregression terms. The level-one intercept terms had an average mean squared error of 0.355, while the level-one growth terms had an average mean squared error of 0.153; finally, the level-one autoregression had an average mean squared error of 0.189. The person-level ability estimates, θ, had the largest average mean squared error of any parameter, with a value of 1.120. The last level-one term, the variance, σ^2, had an average mean squared error of 0.1092. Unlike the results for the mean relative bias, this was roughly comparable to the average mean squared error of the level-two growth variance, 0.0092. The average mean squared error for the average level-two autoregression term γ_3 was 0.0894, while that statistic was 0.0621 for the associated variance term T_{33}. More detail on the mean squared error may be found in Table 5, below.

Table 5. Mean squared errors across conditions.

			Level-one regression terms				Variance terms			Level-two means	
N	φ	Prior distribution	α	β	Φ	θ	σ^2	T_{22}	T_{33}	γ_2	γ_3
100	0	Non-informative	0.402	0.137	0.235	1.131	0.007	0.010	0.063	0.021	0.048
		Weakly informative	0.426	0.151	0.227	1.837	0.014	0.006	0.024	0.029	0.039
		Strongly informative	0.425	0.160	0.206	1.051	0.005	0.017	0.061	0.020	0.022
	0.3	Non-informative	0.387	0.135	0.223	1.246	0.009	0.011	0.066	0.017	0.084
		Weakly informative	0.229	0.240	0.190	0.727	0.156	0.008	0.058	0.008	0.021
		Strongly informative	0.232	0.207	0.213	0.661	0.218	0.009	0.056	0.007	0.014
	0.8	Non-informative	0.363	0.134	0.149	1.441	0.019	0.011	0.071	0.015	0.201
		Weakly informative	0.214	0.215	0.168	0.602	0.419	0.006	0.071	0.004	0.085
		Strongly informative	0.214	0.187	0.190	0.526	0.571	0.004	0.071	0.002	0.078
500	0	Non-informative	0.371	0.118	0.187	0.776	0.045	0.008	0.051	0.012	0.013
		Weakly informative	0.406	0.131	0.208	1.102	0.002	0.012	0.059	0.018	0.025
		Strongly informative	0.407	0.130	0.198	1.012	0.001	0.012	0.057	0.017	0.018
	0.3	Non-informative	0.367	0.120	0.173	0.727	0.157	0.008	0.058	0.008	0.021
		Weakly informative	0.402	0.129	0.218	1.390	0.004	0.011	0.066	0.019	0.076
		Strongly informative	0.394	0.127	0.203	1.243	0.001	0.011	0.064	0.018	0.062
	0.8	Non-informative	0.357	0.116	0.124	0.602	0.419	0.006	0.071	0.004	0.085
		Weakly informative	0.394	0.126	0.201	1.941	0.012	0.008	0.073	0.020	0.328
		Strongly informative	0.252	0.294	0.160	1.947	0.013	0.008	0.073	0.020	0.319
1000	0	Non-informative	0.368	0.116	0.179	0.713	0.063	0.008	0.049	0.010	0.008
		Weakly informative	0.267	0.304	0.163	1.010	0.001	0.012	0.057	0.017	0.018
		Strongly informative	0.404	0.128	0.196	1.013	0.001	0.012	0.057	0.016	0.016
	0.3	Non-informative	0.367	0.117	0.168	0.661	0.218	0.009	0.056	0.007	0.014
		Weakly informative	0.397	0.129	0.214	1.414	0.003	0.011	0.064	0.020	0.078
		Strongly informative	0.396	0.127	0.201	1.285	0.001	0.011	0.064	0.018	0.059
	0.8	Non-informative	0.361	0.108	0.123	0.526	0.571	0.004	0.071	0.002	0.078
		Weakly informative	0.388	0.121	0.197	1.945	0.013	0.008	0.073	0.020	0.319
		Strongly informative	0.384	0.121	0.181	1.728	0.006	0.007	0.073	0.018	0.284

Credible interval coverage rates

The credible interval coverage rates are the final results presented for study one. This should provide evidence addressing the fourth research question, along with the fifth and final research question. Once again, the fourth research question addressed the credible interval coverage rates for the mean and variance of the level-two latent growth, while the final research question evaluated general parameter recovery. Generally, the 95% credible interval covered the generating parameters less than 90% of the time, on average, regardless of the parameter or the simulating condition.

The average credible interval coverage rate for the level-two growth parameter was 18.09%. The credible interval coverage for the non-informative, weakly informative, and strongly informative conditions was 33.03%, 11.68%, and 9.567%, respectively. Coverage rates increased as the autoregressive parameter φ increased; with φ equal to 0, 0.3, and 0.8, coverage rates were 8.597%, 10.64%, and 35.04%, respectively. Finally, the coverage rates actually decreased as sample size increased. Sample sizes of 100, 500, and 1,000 had corresponding coverage rates of 36.56%, 9.881%, and 7.831%, respectively.

There was a roughly similar pattern of results for the variance of the growth term parameter, T_{22}. As prior information increased from non-informative to weakly informative to strongly informative, credible interval coverage decreased from 48.74% to 20.82% to 14.05%. Also, as the sample size increased from 100 to 500 to 1000, the coverage rate decreased from 58.63% to 14.62% to 10.35%. Unlike the

previous result, however, the coverage rate did not appear to vary systematically with the autoregression value φ; as the average autoregression went from 0 to 0.3 to 0.8, the credible interval coverage rate went from 22.57% to 19.62% to 41.41%.

To finish addressing the fifth and final research question of the first study, which addressed general parameter recovery. the credible interval coverage rates for the remaining parameters are reported. The coverage rates were the poorest for T_{33}; the 95% credible intervals included the true values only 1.81% of the time. At level one, the true α, β, and φ parameters were included in their respective 95% credible intervals 80.78%, 73.15%, and 53.43% of the time, respectively. The true ability estimates θ were included in their 95% credible interval 59.99% of the time, while the variance σ^2 was included in its 95% credible interval 45.13% of the time. Finally, the true average autoregression value γ_3 was included in its 95% credible interval 24.72% of the time. More information on credible interval coverage is provided in table 6, below.

Table 6. Credible interval coverage rate percentages by condition.

			Level-one regression terms				Variance terms			Level-two means	
N	φ	Prior Distribution	α	β	φ	Θ	σ^2	T_{22}	T_{33}	γ_2	γ_3
100	0	Non-informative	80.62	73.29	59.06	53.41	97.22	68.86	0	36.71	65.57
		Weakly informative	76.39	69.04	56.33	44.39	92.31	77.23	48.88	20.54	73.76
		Strongly informative	76.41	61.62	61.82	48.91	93.75	34.38	0	15.62	84.38
	0.3	Non-informative	82.95	75.05	56.95	58.60	99.24	73.86	0	46.97	48.86
		Weakly informative	80.49	75.94	74.28	79.82	0.17	42.31	0	21.42	46.46
		Strongly informative	78.63	77.71	75.48	82.49	0	8.18	0	2.83	30.82
	0.8	Non-informative	85.14	77.45	56.73	66.57	98.30	75	0	54.55	5.68
		Weakly informative	80.99	73.32	71.86	91.98	0.69	65.52	0	62.76	1.38
		Strongly informative	77.93	77.60	73.38	93.62	0	82.35	0	67.65	0
500	0	Non-informative	87.83	79.9	65.04	67.89	10.14	19.93	0	4.38	60.14
		Weakly informative	77.13	68.20	53.31	43.21	79.50	0.12	0	0.12	22.29
		Strongly informative	79.35	69.98	57.04	48.09	96.53	0	0	0	43.20
	0.3	Non-informative	89.96	81.60	59.16	79.82	0.17	42.31	0	21.42	46.46
		Weakly informative	76.34	69.25	41.83	41.09	70.76	0.26	0	0.26	0.26
		Strongly informative	79.39	71.34	46.53	46.22	98.48	1.52	0	0	4.55
	0.8	Non-informative	92.41	86.06	47.69	91.98	0.69	65.52	0	62.76	1.38
		Weakly informative	74.33	70.14	24.05	36.77	13.73	1.96	0	0	0
		Strongly informative	72.42	46.01	44.32	36.68	0	0	0	0	0
1000	0	Non-informative	88.88	80.86	65.77	70.86	0	2.64	0	0	47.03
		Weakly informative	75	67.82	63.56	48.13	96.53	0	0	0	43.20
		Strongly informative	79.42	69.72	55.34	47.96	95.95	0	0	0	11.19
	0.3	Non-informative	90.56	81.75	59.51	82.49	0	8.18	0	2.83	30.82
		Weakly informative	76.57	68.71	42.75	40.77	52.79	0	0	0	0
		Strongly informative	78.82	70.82	43.93	45.76	97.83	0	0	0	0
	0.8	Non-informative	92.83	88.70	47.35	93.62	0	82.35	0	67.65	0
		Weakly informative	73.68	70.29	18.78	36.70	0	0	0	0	0
		Strongly informative	76.72	72.85	20.77	41.86	23.81	0	0	0	0

Methods for study two

Participants and Inclusion Criteria

As previously mentioned, the data for this empirical analysis was drawn from the one used in Baams et al., 2015. That study was largely intended to evaluate suicidality among youth who were gay, lesbian, or bisexual, aged 15 to 20. To a large extent, community-based agencies recruited most of these individuals in New York City, San Francisco, and Tucson. The first wave of the data collection, in 2011, was followed by three separate waves, each 9 months later. Aggregate demographic statistics for these individuals are provided in the original study (Baams et al., 2015) and are omitted here because only a subset of the original sample was analyzed.

The inclusion criteria for the empirical analysis were twofold. To avoid potential nesting effects from the recruitment city, only participants recruited in New York City were included. New York City had the largest number of participants, which was thought to be necessary for convergence. Also, only complete cases were analyzed; in other words, participants had to have completely filled out the perceived burdensomeness scale in all four waves of data collection. Future research may better incorporate the missing data by including missing variables in the Gibbs sampler (see, for example, Gefland, Hills, Racine-Poon, and Smith, 1990). A total of 440 participants in New York City were in the dataset, however only 220 of them had complete data for the perceived burdensomeness variable. As such, 220 participants were included in the empirical analysis.

Measures

The only measure included in the empirical analysis was the perceived burdensomeness subscale of the Interpersonal Needs Questionnaire (Van Orden et al., 2008). Seven questions assessed the extent to which participants felt that they were a burden on others. For example, participants responded to the prompt "These days, I think the people in my life wish they could be rid of me". Participants responded on a seven-point scale, with a choice of one corresponding to "not at all true for me", a seven corresponding to "very true for me", and a four corresponding to "somewhat true for me". Baams et al., (2015) reported a high internal consistency statistic for this measure on their full sample ($\alpha = 0.88$).

Analysis plan

For confidentiality reasons, background characteristics, including participant age, ethnicity, gender, and sexual orientation were not provided. Without this information it is difficult to theoretically justify a conditional latent growth model. The empirical model did not regress the growth or initial level of perceived burdensomeness on any other variables because these critical background characteristics were not provided. Instead, an unconditional second-order latent growth model was estimated. Also, since the main research question involved the stability of perceived burdensomeness, an autoregressive parameter was estimated, and the Kalman filter was used to expedite the convergence of the model. Finally, while the original data was on a seven-point scale, the estimation method as described in this document has only dealt with dichotomous data. As such, scores on

each item were dichotomized so that participants who received a score of four or higher on an item, which corresponded with at least somewhat agreeing with the prompt, were differentiated from participants who received a score of lower than a four.

The estimation method largely followed the methods in the simulation study. The Kalman filter, along with other samplers, was used to attempt to obtain posterior distributions of person-level trait estimates, along with the estimated stability, or autoregression, for each participant, along with all the other parameters described in the methods for the first study. Also, similar to study 1, 10,000 iterations of the Gibbs sampler were obtained after 2,000 burn-in iterations. Convergence was again defined as a Gelman-Rubin statistic less than 1.1 for every parameter. Since a nonconvergent model is a more severe issue for a real data analysis, an extra 10,000 iterations of the Gibbs sampler was planned in case convergence failed with the initial iterations. It was also planned that statistics for these parameters would only be reported if convergence was achieved, either with the first or second 10,000 iterations. After this, an additional 20,000 replications were run to see if convergence would be achieved.

Results for study two

Convergence

The model failed to converge after the first 10,000 sampling iterations, which followed an initial 2,000 burn-in iterations. After an extra 10,000 iterations, the model still failed to converge. Finally, after 40,000 total sampling iterations, the model failed to converge. Results are reported for 20,000 sampling iterations, since 40,000 iterations may be more computing time than is reasonable for applied research.

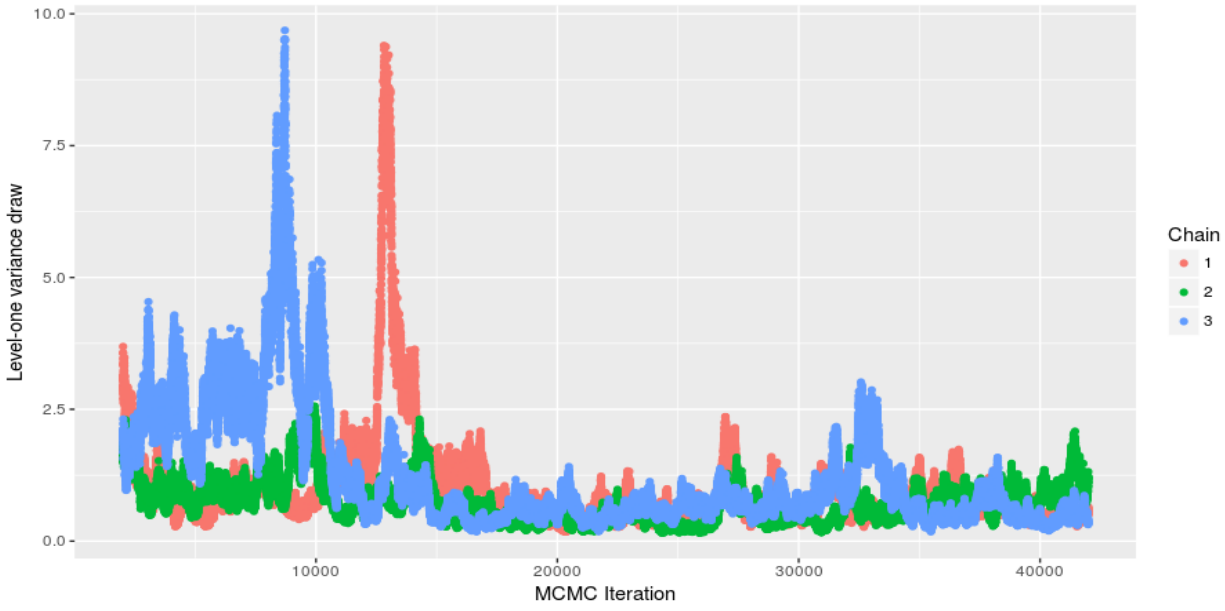
Overall, the Gelman-Rubin statistic, (also known as R-hat) indicated a lack of convergence across all parameters. At the first level, the minimum R-hat for the *a* parameter was 1.913, although six out of the seven *b* parameter R-hat statistics were less than 1.1. The R-hat statistic for the level-one variance, σ^2, was 2.713, and 56.1% of theta parameters had R-Hat statistics greater than 1.1. The last level-one parameters, the level-one regression weights (β) had a mean R-hat statistic of 1.19, and 34.39% of the level-one regression weights' R-hat statistics were greater than 1.1. At level two, the variance matrix **T** mostly had convergent estimates, with the exception of T_{22}, the element which estimated the variance of participant growth. Finally, neither of the level-two means, γ, converged, with R-hat statistics of 1.147 and 1.789 for the level-two growth and autoregression means, respectively.

Because the model failed to converge with 40,000 iterations, median posterior statistics are not reported. Instead, the trace plot of the level-one variance (σ^2) is provided to demonstrate potential convergence problems.

Traceplot

The traceplot for the level-one variance is provided below. The traceplot appears to indicate a great deal of autoregression in the Markov chains for the level-one variance parameter. To investigate this, within chains, the level-one variance for iteration *i* was correlated with the level-one variance for iteration *i* plus *k,* where *k* is the order of the autoregression reported next. The first-order autoregression for the level-one variance was estimated at 0.992, while the tenth-order autoregression was estimated at 0.979. Finally, the twentieth-order autoregression was estimated at 0.969. Additionally, the traceplot shows that the third and first Markov chains obtained draws for the level-one variance were well over 2.5 for several thousand iterations in a row. For the sake of comparison, the frequentist estimate of the level-one variance was 0.09916.

Figure 7. The traceplot of the level-one variance for the empirical analysis. Created using ggplot2 (Wickham, 2009).



Discussion

Convergence rates

The first research question evaluated the convergence rates across all the conditions of the simulation study. Overall, convergence rates for this analysis method were disappointing. One of the goals of this method of analysis was to improve the convergence rates found in Murphy, et al., (2011), which had, at worst, convergence rates of about 57%. In the present work, certain conditions had convergence rates of less than ten percent. With that being said, convergence rates improved with a larger sample size and with smaller autoregressive parameters, a pattern similar to Murphy, et al. (2011).

To be fair, the convergence criterion in Murphy, et al. (2011) is not directly comparable to the criterion used in the present work. In fact, the convergence criteria from these two works are quite distinct conceptually. Recall that checking for convergence in that Murphy, et al. (2011), involved iteratively updating proposed model estimates and checking a discrepancy function. In the present work, the Gelman-Rubin statistic was used to determine convergence (Gelman and Rubin, 1992). Convergence is achieved with the Gelman-Rubin statistic when the ratio of variance between and within independent Markov chains reaches a certain threshold. In other words, the model is said to have converged when the chains have sufficiently "mixed". In that case, the samples from the posterior distribution are roughly comparable from all three Markov chains. Because these two convergence

criteria are distinct, both theoretically and mathematically, they are not directly comparable to each other.

Despite this, it remains somewhat unusual that the convergence rates in this work degraded with higher autoregressive parameters. Certainly, if the Kalman filter was not used, this sort of result would be expected. In that case, the theta parameter for a particular person at time *t* would depend on that persons' theta parameter at time *t-1*, to a degree proportional to the autoregressive parameter. In that situation, a higher autoregressive parameter would ultimately increase the between-chain variance and increase the Gelman-Rubin statistic. However, using the Kalman filter ensures that all the theta parameters are sampled at the same time, removing this dependency and, in theory, reducing the Gelman-Rubin statistic.

One possible explanation is that the variance-covariance matrix **T** may be more often computationally singular when sampling a dataset with a higher autoregressive parameter. In this study, there were sometimes iterations in which **T** was computationally singular, making it impossible to sample γ_2 and γ_3. When this happened, that iterations' γ_2 and γ_3 draws were set to be equal to the previous iterations' draws. This would certainly increase the between-chain variance for the γ_2 and γ_3 parameters in cases in which **T** was computationally singular for several iterations in a row. Ultimately, in those cases, one would expect a higher Gelman-Rubin statistic and lower convergence rates.

The main reason one might expect $\mathbf{T}$ to be computationally singular more often for higher mean autoregressive conditions is that every level-one autoregressive parameter is restricted to be between negative and positive one. The closer the person-level autoregressive parameters are to one, the less variance there may be in those estimates, due to this restriction of range. This reduced variance, in turn, may result in level-two variance estimates that are close to zero, resulting in a computationally singular $\mathbf{T}$ matrix. In the future, allowing the autoregressive parameter to be greater than one in absolute value may be reasonable for educational research. This restriction is in place mainly since it allows for long-term forecasting (Hamilton, 1994), which may be unnecessary for educational research. Additionally, removing this restriction would allow the autoregressive parameters to be sampled as random draws from a Normal distribution, rather than a truncated Normal distribution, which may be less computationally expensive.

In general, the low convergence rates might be improved by thinning the Markov chains. In MCMC estimation, thinning involves only including every k^{th} sample in the final analysis (Link and Eaton, 2011). This subsampling helps to ensure that each sample in a chain is nearly independent of the previous sample in that chain. This, in turn, should reduce the between-chain variance compared to the within-chain variance, ultimately improving the Gelman-Rubin statistic and therefore improving convergence. Unfortunately, implementing thinning was not computationally feasible for this project with the proposed number of conditions and

replications per condition. It is suggested that future simulation studies using this estimation method have fewer conditions to incorporate thinning.

However, the findings from the empirical study suggest that thinning the chains may have limited utility, particularly for the level-one variance parameter. Even if only one out of every twenty iterations of the Markov chain were kept, the autoregression would be 0.969, indicating a great deal of autoregression. A higher autoregression in this case would tend to increase the ratio of between-chain to within-chain variance, ultimately reducing convergence rates. It is unfortunate that the applied research question of study 2 cannot be answered due to this lack of convergence, however the autoregression within the chains elucidates issues with the estimation method.

This may suggest that using only a Gibbs sampler with the Kalman filter is an inadequate estimation method for second-order autoregressive latent growth models. Instead, a more specialized Metropolis-Hastings algorithm may be needed (Gamerman and Lopes, 2006). Rather than sampling the level-one variance directly from its conditional distribution, the Metropolis-Hastings algorithm would involve the use of a proposal distribution, which may be adjusted to ultimately reduce autoregresison in the Markov chains. However, the incorporation of this algorithm would require a fair amount of mathematical and theoretical work, which may also depend somewhat on the research area. Future research is needed to more fully incorporate the Metropolis-Hastings algorithm to estimate second-order autoregressive latent growth models.

When considering the Metropolis-Hastings algorithm, one might also consider using Stan (Carpenter et al., 2017) to estimate this class of model. This estimation software uses a high-dimensional Metropolis-Hastings proposal distribution to reduce correlations between parameter draws (Neal, 2011). However, to my knowledge, the Kalman filter cannot be used in Stan. Whereas the Kalman filter provides a specific strategy for Gibbs sampling, Stan necessarily uses a particular Metropolis-Hastings proposal distribution. The latter is mathematically distinct from pure Gibbs sampling. As such, Stan cannot use the Kalman filter to explicitly account for the structure of autocorrelated time series data. However, since Stan is intended to reduce correlations between parameter draws, possibly the software would improve convergence regardless.

Finally, it is noted that this study used a rather strict definition of convergence with respect to the Gelman-Rubin statistic. Specifically, the model was only said to have converged when the Gelman-Rubin statistic was below 1.1. Other analysts (Gelman and Rubin, 1992; Fox and Glas, 2001; Gelman et al., 2014) have suggested a value of 1.2 instead. The stricter threshold was chosen since the present study attempted to improve convergence under rigorous conditions. Particularly, a greater convergence rate with a strict threshold would have provided strong evidence for convergence. One would expect more models to converge if 1.2 was the threshold instead of 1.1, however all the issues in the above paragraphs would remain.

Recall that the first research question of the first study asked what the convergence rates are for a second-order autoregressive latent growth model when estimated using a Gibbs sampler with the Kalman filter. Generally, the evidence suggests that these convergence rates are much lower than those found in Murphy et al. (2011), and even lower when strongly and weakly informative priors are used for item parameters. This last finding contradicts the original hypothesis that more strongly informative priors would improve convergence rates; this may be due to a potential scaling problem, discussed next.

Scaling problem

Both the informative priors and the data generation process were based on models that were identified differently from the estimated model. Specifically, the item parameters taken for the informative priors were based on a model in which the average ability estimate was zero, with a standard deviation of one. In contrast, with the second-order latent growth model, the ability estimates of only the first time point had a fixed average of zero, with a fixed standard deviation of one. For later time points, the true average ability estimate increased due to the growth parameter. As such, the scaling of the item parameters differed. Also, in the estimation, the discrimination and difficulty parameter of the first item were always set to one and zero, respectively, however data were not generated this way, instead following Table 1 directly.

There is another scaling problem for the data generation process. Following the data generation, first the ability estimates for each person at each time point

were generated. Next, the probability of a correct response for a specific item at a specific time point was set to be proportional to the discrimination parameter times that ability estimate minus the difficulty parameter. However, again, after the first time point the average ability estimate was not equal to one. While the item parameters did not change in the data generation process over time, both the mean and the variance of the ability estimates changed over time. Therefore, for the later time points, the true item parameters are scaled differently from those provided in Table 1, which were estimated with ability estimates that had a mean of zero and a variance of one. Finally, when the model was finally estimated, it is expected that the overall average ability estimate was greater than zero, with a variance higher than one. These problems ultimately resulted in mean relative biases for item parameters that were too extreme to be usefully reported. Specifically, item difficulty parameters would tend to be underestimated, since the theta values were, on average, shifted upward. Possibly for this reason, item parameter recovery was degraded for the difficulty parameters when more informative priors were used, as shown in Table 4. While it is encouraging that item discrimination parameter recovery appeared to improve when more informative priors were used, overall the differences by condition for both discrimination and difficulty were quite small, and should be interpreted with caution given the differential convergence rates of the conditions.

Conditions with sample sizes of 500 or 1000 did not converge as often when weakly or strongly informative priors were used. Mathematically, it appears that

item parameters would be unlikely to converge when the prior strongly favors a value that is disparate from the likelihood function. Also, these uncertain item parameters may explain why individual-level ability estimates were generally more poorly estimated using more strongly informative priors. Recall that the final research question assessed general parameter recovery across conditions. As shown in Tables 3, 5, and 6, above, mean relative bias, mean squared error, and the credible interval coverage rates for the theta estimates were generally degraded when more strongly informative priors were used. Again, it mathematically feasible that differently scaled item parameters may have led to this reduction in parameter recovery for the theta estimates. Of course, the theta estimates are just one set of parameters in the model; a more complete discussion of parameter recovery is provided next.

Parameter recovery

To finish addressing the final research question of the first study, the general parameter recovery of the level-one regression weights and the level-one variance is addressed here. Recovery of the theta estimates and item response theory estimates was addressed in the previous section, and the recovery of the level-two parameters is the content of the remaining research questions for the first study. Overall, it appears from table 3 that the variance term σ^2 was overestimated, while the remaining variance terms T_{22} and T_{33} were underestimated.

The underestimation of T_{33}, in particular, may be due to some alterations to the estimation method. It was originally proposed that the Kalman filter would use

person-specific autoregression parameters. To be precise, the Markov chain would use each simulee's φ value in Equation 36 to ultimately sample θ, the ability estimates. However, preliminary findings showed even more extreme convergence problems than those presently reported. In fact, convergence rates actually decreased as the sample size increased. To attempt to address the convergence problem found in the study, instead the value λ_3 was used in Equation 36. This did improve convergence rates dramatically, however with this change the individual-level autoregression values are not reflected in the likelihood function of the ability values. This may be the main reason why the variance term, T_{33}, was underestimated. Also with this change, we would expect person-level φ values to be estimated as randomly varying around λ_3, causing severe bias overall.

With an underestimated level-two variance term, and a fixed total variance, it then follows that the variance term σ^2 would be overestimated, generally. Furthermore, with these large biases in place, it is unsurprising that the credible interval coverage rates for the variance terms were generally quite poor. As suggested previously, future research may attempt to address this issue by allowing the individual-level autoregression terms to be greater than one in absolute value. If convergence problems persist, then individual-level autoregression parameters would not be recommended for second-order latent growth models.

The level-one regression terms were also generally underestimated. As suggested before, the cause for the bias in φ terms may be due to the change in the estimation method, using λ_3 instead of individual-level φ terms in Equation 36,

above. The scaling problem may cause the remaining regression terms, α and ß, to be underestimated. The pool of simulees generally has an average ability estimate greater than zero, making the item difficulty terms lower than they should be. If the items appear to be very easy, answering more items correctly would not indicate as much person-level growth as we would expect. Similarly, not answering very many items correctly in the first wave of data collection would indicate a very low ability estimate, which in turn would cause α to be similarly underestimated.

Despite these severe problems with parameter recovery, the correlations reported in Table 4, above, are encouraging. The correlations, typically greater than 0.9, suggest that if the scaling problems are corrected, the estimation procedure may improve. More statistical mathematics is needed to more properly relate a test calibration to panel data.

The final research question of the first study asked the conditions under which parameter recovery is impacted by using more strongly informative priors. Evidence suggests that parameter recovery is generally quite poor, regardless of the priors used for item parameters. Future research may improve parameter recovery by fixing the scaling problem and allowing autoregression values to be greater than one in absolute value.

Estimation of Latent Growth, by Condition

The second, third, and fourth research questions of the first study investigated the mean relative bias, mean squared error, and credible interval

coverage rates of the λ_2 and T_{22} terms, respectively. Once again, λ_2 refers to the level-two average latent growth term, while T_{22} refers to the level-two variance of simulee growth. Across these disparate outcomes, a clear pattern emerges. The level-two growth term λ_2 is more poorly estimated than T_{22}, and estimation of both is degraded when more informative priors are used. These findings are unsurprising given the underestimation of item difficulty and the scaling problem. If items are estimated to be easier than the generating values, it makes sense for growth to be underestimated. However, problems with convergence confound this finding. Since more strongly informative priors generally did not converge as often as the non-informative cells, the mean relative bias, mean squared error, and credible interval coverage of the level-two terms are more poorly estimated for more strongly informative conditions. Therefore, these relatively small differences in mean relative bias should be interpreted with caution. At most, they provide additional modest evidence for a scaling problem.

The convergence problems render the effects of varying the sample size and the autoregression uninterpretable. Sample size had a strong impact on model convergence, particularly when the autoregression was high. Therefore, any differences in mean relative bias by sample size or autoregression may be due to the estimation process, or they may be due to unbalanced cells in Table 3. Indeed, one would expect the larger relative bias for smaller sample sizes to at least partly arise from convergence issues. For this reason, although results are presented showing

the average mean relative bias across these conditions, they are not interpreted here.

To summarize, the estimation of the level-two growth terms was degraded when more informative priors were used. This provides additional evidence for a scaling problem. The relationships between sample size, autoregression, and model convergence confound the effects of those conditions on the estimation of the level-two growth terms.

Limitations

The difficulty of achieving convergence, along with extreme computational difficulties, limits these findings. Certain cells in the above tables were estimated based on many more or many fewer replications. For this reason, many results should be interpreted with caution; the tabulated statistics may themselves be subject to differential simulation error.

The scaling problem with the item parameters presents another issue. Also, one of the advantages of second-order latent growth modeling is that item drift can be assessed over time, and the psychometric properties of different items at different time points may be evaluated. The present work is limited by simulating a single set of items which do not drift over time. Future research may extend this work by simulating item drift or including particularly easy or difficult time points at particular time points. Mainly, however, the differential scaling of the item parameters may have impacted parameter recovery across all conditions. More

statistical mathematics may be needed to devise informative priors for item responses that have the same scale as the estimated model.

The first study does not simulate any missing data, which is typically unrealistic for longitudinal datasets. Most large longitudinal datasets have participant attrition over time; also, participants may not be available for particular testing sessions. This was, in fact, the case in study two, in which only complete cases were included. In theory, the Gibbs sampler as described here may be able to accommodate these missing data structures. That is, the sampler may estimate participant ability under missing data by using Equation 36, above, and use that term for the next iteration of the sampler. This relatively straightforward extension may be a fruitful topic for further research.

The simulations are also somewhat contrived in that it is assumed that every participant is measured at equal intervals. Occasionally participants are tested at different time points due to practical concerns. More typically in longitudinal datasets, the spacing between testing occasions is somewhat staggered from participant to participant. Future research may address this practical issue by simulating slightly different values of the time variable t in Equation 36, above, for each participant.

Summary and Recommendations

This research necessitated the use of over a century of computer time. The duration of the computations may be reduced by coding the sampler in C++

exclusively, rather than using R and RCpp. Also, in future research, computing time may be further reduced by allowing the autoregression terms to be greater than one in absolute value, rather than using rejection sampling.

Attempting to estimate individual-level autoregression terms in Equation 36, above, resulted in severe convergence problems, which were attenuated, but not eliminated, by using λ_3 instead of φ in the Kalman filter. Convergence problems, which made the evaluation of the remaining research questions difficult, may be addressed in two ways. First, the identification may be corrected so that the informative priors and generating data are on the same scale as the estimated model. Secondly, the Metropolis-Hastings algorithm, instead of the Gibbs sampler, may be used to estimate the level-one variance, along with other terms that exhibit a great deal of within-chain autoregression. This would tend to reduce the ratio of between-chain to within-chain variance, improving convergence.

The generating model, the estimating model, and the model used to form informative priors were all identified differently. In the generating model, all the IRT parameters were set to be equal to those found in Table 1, but ability estimates generally had a mean greater than zero and a variance greater than one. In the estimating model, the first item had a fixed difficulty and discrimination of zero and one, different from the first item in Table 1. In the model used to form informative priors, the IRT parameters were again set to those found in Table 1, but ability estimates had a mean and variance of zero and one, respectively. Results in Table 4 suggest that parameter recovery may dramatically improve once this scaling issue

is resolved. Possibly because of this scaling problem, the estimation of level-two latent growth became worse when more informative priors were used. Despite these problems with estimation, the theoretical justification for using informative priors to estimate second-order latent growth models has not changed. In Item Response Theory, item calibration is said to be person-free (Lord, 1952), corresponding to the Bayesian assumption of exchangeability (de Finnetti, 1974; Lindley & Novick, 1981; Rubin, 1978). It is hoped that future researchers will rectify this scaling problem to use more informative priors for latent growth modeling.

Works Cited

Baams, L., Grossman, A., & Russell, S. T. (2015). Minority Stress and Mechanisms of Risk for Depression and Suicidal Ideation Among Lesbian, Gay, and Bisexual Youth. *Developmental Psychology, 51*(5), 688-696.

Barnes, G. M., Reifman, A. S., Farrell, M. P., & Dintcheff, B. A. (2000). The effects of parenting on the development of adolescent alcohol misuse: a Six-Wave latent growth model. *Journal of Marriage and Family, 62*(1), 175-186.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective* (Vol. 467). John Wiley & Sons.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2008). *Time series analysis: Forecasting and control* (4th ed.). New York, NY: Wiley.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics, 7*(4), 434-455.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software* 76(1).

Catts, H. W., Bridges, M. S., Little, T. D., & Tomblin, J. B. (2008). Reading achievement growth in children with language impairments. Journal of Speech, Language, and Hearing Research, 51(6), 1569-1579.

Chan, D., & Schmitt, N. (2000). Interindividual differences in intraindividual

changes in proactivity during organizational entry: a latent growth modeling approach to understanding newcomer adaptation. *Journal of Applied Psychology, 85*(2), 190.

Chatfield, C. (2004). *The analysis of time series* (6th ed.). Boca Raton, FL: Chapman-Hall/CRC Press.

Colder, C. R., et al., (2001). Identifying trajectories of adolescent smoking: an application of latent growth mixture modeling. *Health Psychology*, *20*(2), 127.

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research, 38*(4), 529-569.

de Finnetti, B. (1974). *Theory of probability.* New York: Wiley.

Duncan, S. C., Duncan, T. E., Biglan, A., & Ary, D. (1998). Contributions of the social context to the development of adolescent substance use: a multivariate latent growth modeling approach. *Drug and alcohol dependence, 50*(1), 57-71.

Elliott, M. R., Gallo, J. J., Ten Have, T. R., Bogner, H. R., & Katz, I. R. (2005). Using a Bayesian latent growth curve model to identify trajectories of positive affect and negative events following myocardial infarction. *Biostatistics (Oxford, England)*, *6*(1), 119.

Ferrer, E., Balluerka, N., & Widaman, K. F. (2008). Factorial invariance and the specification of second-order growth models. *Methodology, 4*, 22–36.

Fleming, C. B., Mason, W. A., Mazza, J. J., Abbott, R. D., & Catalano, R. F. (2008).

Latent growth modeling of the relationship between depressive symptoms and substance use during adolescence. *Psychology of Addictive Behaviors, 22*(2), 186.

Fox, J., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66*(2), 271-288.

Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference.* CRC Press.

Geiser, C., Keller, B. T., & Lockhart, G. (2013). First-versus second-order latent growth curve models: some insights from latent state-trait theory. *Structural equation modeling: a multidisciplinary journal*, *20*(3), 479-503.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). Boca Raton, FL, USA: Chapman & Hall/CRC.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevelhierarchical models* (Vol. 1). Cambridge: Cambridge University Press.

Gelman, A. and D. B. Rubin (1992) *Inference from iterative simulation using multiple sequences (with discussion)*. Statistical Science, 7:457-511.

Granger, C. W. J., & Morris, M. J. (1976). *Time Series Modeling and Interpretation.* Journal of the Royal Statistical Society, Series A (General), 139(2), 246-257.

Grimm, K. J., & Ram, N. (2009). Nonlinear growth models in M plus and SAS. *Structural Equation Modeling, 16*(4), 676-701.

Grimm, K. J., Ram, N., & Hamagami, F. (2011). Nonlinear growth curves in developmental research. *Child development, 82*(5), 1357-1371.

Hamilton, J. D. (1994). *Time series analysis* (Vol. 2). Princeton: Princeton university press.

Hancock, G. R., Harring, J. R., & Lawrence, F. R. (2013). Using latent growth modeling to evaluate longitudinal change. In G. R. Hancock & R. O. Mueller (Eds.), *Structural Equation Modeling: A Second Course (2nd ed) (pp. 309-341).* Charlotte, NC: Information Age Publishing.

Hancock, G. R., Kuo, W. L., & Lawrence, F. R. (2001). An illustration of second-order latent growth models. *Structural Equation Modeling, 8*(3), 470-489.

Harring, J. R., Kohli, N., Silverman, R. D., & Speece, D. L. (2012). A second-order conditionally linear mixed effects model with observed and latent variable covariates. *Structural equation modeling: a multidisciplinary journal, 19*(1), 118-136.

Jordan, N. C., Hanich, L. B., & Kaplan, D. (2003). A longitudinal study of

mathematical competencies in children with specific mathematics difficulties versus children with comorbid mathematics and reading difficulties. *Child development, 74*(3), 834-850.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of basic Engineering, 82*(1), 35-45.

Kline, R. B. (2011). Convergence of Structural Equation Modeling and Multilevel Modeling. In Williams, M., & Vogt, W. P. *Innovation in Social Research Methods* (pp. 562-589). Thousand Oaks, CA: Sage Publications Inc.

Kruschke, J. K. (2011). *Doing Bayesian Data Analysis: A Tutorial with R and BUGS.* Burlington, MA: Academic Press.

Kwok, O. M., West, S. G., & Green, S. B. (2007). The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: A Monte Carlo study. *Multivariate Behavioral Research, 42*(3), 557-592.

Leite, W. L. (2007). A Comparison of Latent Growth Models for Constructs Measured by Multiple Items. *Structural Equation Modeling, 14*(4): 581-610.

Lindley, D., V., & Novick, M. R. (1981). The role of exchangeability in inference. *Annals of statistics, 9*, 45-58.

Link, W. A., & Eaton, M. J. (2011). On thinning of chains in MCMC. *Methods in*

Ecology and Evolution, 3(1), 112-115.

Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods, 18*(3), 285.

Lord, F. (1952). A theory of test scores. Psychometric monographs.

Murphy, D. L., Beretvas, S. N., & Pituch, K. A. (2011). The effects of autocorrelation on the curve-of-factors growth model. *Structural Equation Modeling: A Multidisciplinary Journal, 18*(3), 430-448.

Murphy, D. L., & Pituch, K. A. (2009). The performance of multilevel growth curve models under an autoregressive moving average process. *The Journal of Experimental Education, 77*(3), 255-284.

Muthén, B., & Asparouhov, T. (2011). Beyond multilevel regression modeling: Multilevel analysis in a general latent variable framework. Handbook of advanced multilevel analysis, 15-40.

Muthén, L. K., & Muthén, B. O. (2007). Mplus. *User's guide.*

Neal, R. (2011). MCMC using Hamiltonian Dynamics In Brooks, S., Gelman, A., Jones, G.L., & Meng, X.L. (eds.), *Handbook of Markov Chain Monte Carlo*, (pp. 116–162). Chapman & Hall/CRC.

Petris, G., Petrone, S., & Campagnoli, P. (2009). *Dynamic Linear Models with R.* New York, NY: Springer Science+Business Media.

Rizopoulos, D. (2006). ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses, *Journal of Statistical Software, 17* (5), 1-25.

Roberts, G. O., & Smith, A. F. (1994). Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic processes and their applications*, *49*(2), 207-216.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, 34-58.

Sayer, A. G., & Cumsille, P. E. (2001). Second-order latent growth models.

Schumacker, R., & Lomax, R. G. (2010). Structural equation modeling. *NY. Routlege.*

Steyer, R., Ferring, D., & Schmitt, M. J. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment*.

Texas Advanced Computing Center (November 2017). *Stampede2 User Guide.* Retrieved from https://portal.tacc.uteas.edu/user-guides/stampede2.

Van Orden, K. A., Witte, T. K., Gordon, K. H., Bender, T. W., & Joiner, T. E., Jr. (2008). Suicidal desire and the capability for suicide: Tests of the interpersonal-psychological theory of suicidal behavior among adults. *Journal of Consulting and Clinical Psychology, 76*, 72–83.

von Oerzen, T., Hertzog, C., Lindenberger, U., & Ghisletta, P. (2010). The effect of

multiple indicators on the power to detect inter-individual differences in change. *British Journal of Mathematical and Statistical Psychology, 63*, 627-646.

Wickham, H. *ggplot2: Elegant Graphics for Data Analysis.* New York, NY: Springer-Verlag. Wright, B. D. (1967). Sample-free test calibration and person measurement. In Bloom, B. S. (Chair): *Invitational Conference on Testing Problems*. Princeton NJ: Educational Testing Service. Pp. 85-101.

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian Mediation Analysis. *Psychological Methods, 14*, 301-322.